AFRL-RH-WP-TR-2011-0018

NATURALISTIC MODEL OF CAUSAL REASONING: DEVELOPING AN EXPERIENTIAL USER GUIDE (EUG) TO UNDERSTAND FUSION ALGORITHMS AND SIMULATION MODELS

Gary Klein
Shane Mueller
Louise Rasmussen
Applied Research Associates, Inc.
1750 Commerce Center Boulevard, North
Fairborn OH 45324

Robert Hoffman
Florida Institute of Human and Machine Cognition
40 South Alcaniz Street
Pensacola FL 32502

SEPTEMBER 2010
Final Report

Distribution A: Approved for public release; distribution is unlimited.

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2011-0018 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNED//
CRAIG M. STANSIFER
Work Unit Manager
Behavior Modeling Branch

//SIGNED//
DAVID G. HAGSTROM
Anticipate & Influence Behavior Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
23-09-2010	Final	January 2008 - September 2010

4. TITLE AND SUBTITLE

Naturalistic Model of Causal Reasoning: Developing an Experiential User Guide (EUG) to Understand Fusion Algorithms and Simulation Models

5a. CONTRACT NUMBER FA8650-04-D-6546 DO 0013

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER 63231F

6. AUTHOR(S)

[1]Gary Klein, [2]Robert Hoffman, [1]Shane Mueller, [1]Louise Rasmussen

5d. PROJECT NUMBER 2830

5e. TASK NUMBER 05

5f. WORK UNIT NUMBER 28300509

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

[1]Applied Research Associates, Inc.
1750 Commerce Center Boulevard North
Fairborn OH 45324

[2] Florida Institute of Human and Machine Cognition
40 South Alcaniz Street
Pensacola FL 32502

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Anticipate & Influence Behavior Division
Behavior Modeling Branch
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)

711 HPW/RHXB

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-RH-WP-TR-2011-0018

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

88ABW/PA cleared on 23 Feb 2011, 88ABW-2011-0712.

14. ABSTRACT

For this effort a naturalistic study of causal reasoning was performed. One strand of this project examined causal reasoning in military and other contexts. A second strand designed a support concept to help people understand the causal reasoning of computer decision aids they use. For the first strand, we determined that most causal reasoning involve indeterminate situations. We demonstrated that people see a variety of causal reasoning formats in naturalistic settings, ranging from simple, single cause attributions to complex events, decisions, and forces. We further demonstrated that we could manipulate preferences for explanation types by varying features of the target audience for an explanation. We also found cultural differences in preferences for simple vs. complex causal reasoning formats. For the second strand, we developed the concept of an Experiential User Guide (EUG), to enable users of sophisticated decision aids to better understand the causal reasoning embedded in the algorithms. We demonstrated this EUG concept with JCAT (Java Causal Analysis Toolkit), a Bayesian reasoning support tool, and formulated recommendations for similar EUG applications with other types of decision aids. Finally, we formulated recommendations for helping people become more effective in causal reasoning within naturalistic settings.

15. SUBJECT TERMS

Causal Reasoning, Explanations, Cognition, Naturalistic Decision Making, Decision Aids

16. SECURITY CLASSIFICATION OF: UNCLASSIFED			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
c. REPORT	b. ABSTRACT	c. THIS PAGE			Craig M. Stansifer
U	U	U	SAR	76	19b. TELEPHONE NUMBER (include area code) NA

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18

THIS PAGE LEFT INTENTIONALLY BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

The Human Predictive Reasoning for Group Interactions research effort was sponsored by the Air Force Research Laboratory's (AFRL), Sensemaking and Organizational Effectiveness Branch (AFRL/RHXS) under Task Order #13 of the Technology for Agile Combat Support (TACS) contract (FA8650-D-6546). The period of performance for the research effort extended from 9 January 2008 to 23 September 2010. This report documents the results of research activities conducted as part of this Task Order.

ACKNOWLEDGEMENTS

The research reported on here was performed under Subcontract No. TACS-SC-04-144, TO 0013 from Northrop Grumman IT (Prime Contract FA8650-04-D-6546). We would also like to thank Dr. Janet Miller, Cheryl Batchelor, and Craig Stansifer with AFRL/RHXB/RHXS for their encouragement and guidance.

SUMMARY

A naturalistic study of causal reasoning was performed during this research project. One strand of this project was to examine causal reasoning in military types of contexts. The second strand was to design a support concept to help people understand the causal reasoning of computer decision aids they use.

For the first, we determined that most causal reasoning involves indeterminate situations where people will never figure out the “real” cause of an effect – even the concept of a real cause is an oversimplification. Events have a variety of causes, and the complexity of naturalistic settings prevents people from using formal logic to pin down a real cause. We further demonstrated that there is a variety of causal reasoning formats people use in naturalistic settings, ranging from single cause attributions to events, decisions, and forces. These also include lists of these single causes, as well as more complex causal maps showing relationships between single causes. We demonstrated that we could manipulate preferences for explanation types by varying features of the target audience for an explanation. We also found cultural differences in preferences for simple vs. complex causal reasoning formats.

Second, we developed the concept of an Experiential User Guide (EUG), to enable users of sophisticated decision aids to better understand the causal reasoning embedded in the algorithms. We demonstrated this EUG concept with JCAT (Java Causal Analysis Tool), a Bayesian reasoning support tool, and formulated recommendations for similar EUG applications with other types of decision aids.

We also presented some recommendations for helping people become more effective in causal reasoning in naturalistic settings.

1.0 INTRODUCTION

The key issue we investigated in this research effort was the psychological nature of causal reasoning. Causal reasoning is central to many of the macrocognitive functions (Klein et al., 2003) and to the macrocognitive activities of Air Force personnel. It drives decision making – the causal models people hold will determine the way they recognize and categorize situations and the kinds of mental simulation they will perform to evaluate courses of action. It is central to sensemaking – the application of causal reasoning to understand events and to modify their causal models based on what they learn. It is central to replanning – diagnosing why a plan might be going poorly and considering what needs to be altered. It is central to coordination – anticipating how the actions of individuals will affect the activities of the team. And, of course, causal reasoning is central to the process of forming and revising mental models. In many, if not most, cases our mental models hinge upon sets of causal knowledge and beliefs we summon to make sense of events. Therefore, a next step in the evolution of macrocognition is to gain a better understanding of causal reasoning.

Causal reasoning has received enormous attention from different communities – philosophers, scientists, economists, historians, and educators, because of its centrality to the ways we think and make sense of events, the ways we learn from experience, and the ways we codify knowledge. Causal reasoning plays a central role in our mental models about how things work and what will happen if we intervene in different ways. Military leaders depend on causal reasoning to select and evaluate courses of action, to gauge their progress, or explain why they are running into trouble. Physicians depend on causal reasoning when they diagnose their patients.

The purpose of this effort was two-fold. First, we explored the nature of causal reasoning and what counts as a useful and acceptable causal explanation. Special emphasis was placed on studying causal reasoning of Air Force (AF) decision makers. Second, we sought to help decision makers understand the "causal reasoning" of software systems by incorporating into simulations (of an appropriate fidelity) a "guide" that would meaningfully illustrate the simulations inner workings and boundary conditions. This kind of *Experiential User Guide (EUG)* could have great value in itself and could serve as a platform for practicing what we learn about causal reasoning. The two research streams described above interacted, with the causal reasoning research informing and guiding the design of the EUG and the evaluations of the EUG the model of causal reasoning.

Managers rely on causal reasoning to figure out who to blame for failure and who to reward for success, and sometimes get it wrong. For example, (Mlodinow, 2008) describes the case of Sherry Lansing, the former head of Paramount Pictures. Under her leadership, Paramount had its greatest financial success from movies such as *Forrest Gump*, *Braveheart*, and *Titanic*. Then Lansing hit a slump. Paramount's market share decreased over six years, from 11.4% to 10.6%, to 7.4%, to 7.1% to 6.7%. The trend was clear and Lansing was fired. Sure enough, Paramount increased its market share the next year to almost 10% with films like *War of the Worlds* and *The Longest Yard*. That seemed to vindicate the studio's decision to dump Lansing except that these were movies that Lansing had put into production. The causal reasoning had gotten confounded with normal variations in box office returns.

1.1 What Counts as a Cause?

The investigation of causality is usually traced to Aristotle but, for our purposes, the account offered by Hume (1739-1740) is much more in line with our modern notion of physical cause-and-effect, although it has been subject to multiple interpretations and criticisms. In the Humean view, in order for there to be an objective establishment of a cause-effect relation, there is some sort of "necessary connection" between the cause and the effect, which is only one item in Hume's list of criteria.

Based on Hume's analysis and related work, we have identified three primary criteria to establish what counts as a cause: propensity, mutability, and covariation.

1.1.1 Propensity. The propensity criterion is that the proposed cause has to plausibly lead to the effect. This is similar to Hume's notion of necessary connection. A hundred years ago a few medical researchers suggested that mosquitoes somehow caused malaria and yellow fever. They were ridiculed because no one could see how tiny mosquitoes could contain enough venom to sicken and kill grown men. It was not until viruses were identified that the mosquito link was understood (Parker, 2008).

The firing of Sherry Lansing falls in this category. The Board of Directors could explain Paramount's slide by considering that when Lansing started she had a fresh vision but, inevitably, her vision became less fresh over time, and was going to get progressively less successful.

A putative cause has to plausibly result in an effect, and the strength of the cause will depend on the links between it and the effect. The more links, the less plausible. The strength is generally no greater than the weakest plausible link in the chain.

1.1.2 Reversibility. The reversibility criterion (usually referred to as "mutability" in the literature) is that the effect should disappear if the putative cause disappears. Kahneman and Varey (1990) linked this notion to counterfactual reasoning, where we can imagine that the proposed cause did not happen – perhaps the star basketball player missed his last-second shot instead of making it. Then, the 1-point victory would turn into a 1-point defeat. In domains such as sports, last minute events can gain causal prominence because they are easiest to mentally reverse. Kahneman and Varey refer to these as "close counterfactuals." A cause is identified by tracing back from the effect to the nearest plausible candidate in the causal chain. The person responsible is the one whose actions cannot easily be reversed by anyone else.

It is possible to imagine reversals that are not close in time to the effects, but the greater the time lag the more complicated (and uncertain) the causal reasoning. Dörner (1986) has shown that participants in a microworld task struggle to make sense of causal connections as the time delay increases. Other than a simple memory problem, time lags permit intervening factors to tangle up the assessment.

The reversibility criterion lets us distinguish causes from "enabling conditions." If someone lights a match and holds it under a piece of paper and the paper begins to burn, we would say that the match caused the burning. We would not say the oxygen in the room caused the burning. Oxygen is necessary for the paper to burn but it is an enabling condition. We can more readily imagine that the match was not held under the paper than the room was void of oxygen.

1.1.3 Covariation. Covariation is the observed coincidence of causes and effects. This covariation contingency is discovered through statistical regularities rather than propensity or reversibility. If we set up a matrix of cause (present or absent) and effect (present or absent), we would find many observations in the upper left-hand corner, where both are present, and the lower right-hand corner – that the effect is rarely if ever seen if the cause has not occurred. The other diagonal would be sparsely populated – few, if any, cases where the cause occurred but not the effect, and perhaps no cases of the effect without the cause. This criterion is related to the "method of differences," described by Mill (1843) as an experimental design for discovering cause-effect relations. Using covariation, we find the biggest difference in situations where the effect occurred or did not occur, and call that the likely cause. (Mill's method of differences was more an all-or-none – if the cause is eliminated the effect goes away – rather than a magnitude relation.) The persistence of medical authorities in Havana in trying to eradicate yellow fever by controlling the mosquito population was due to the strength of the relationship between the two, even in the absence of a plausible causal story.

In applying these three primary criteria, we need to take a few other considerations into account. Context is one consideration. Einhorn and Hogarth (1986) noted that if we see a hammer strike and shatter a watch crystal, we would say the hammer was the cause of the crystal's destruction. But if the observation took place in a watch factory where the hammer was used to test the crystals, and this was the only crystal that shattered, we would say the crystal must have been flawed. But, if the test hammer shattered crystal after crystal, we might speculate that the hammer force was set too high. In most cases, shorter delays strengthen our confidence in the causal connection but the interpretation of delay is actually a function of our mental model of the causal chain. Thus, if you begin to smoke cigarettes today and then, at a routine health screen tomorrow, you find out you have lung cancer, the delay is too short for us to ascribe your cancer to the fact that you began smoking.

In our research, we found an additional causal criterion was also very important – manipulability. If something is a cause of a certain effect, then by manipulating that potential cause, we can modify or alter the effect in question. The manipulability criterion runs into problems of circularity, but on a psychological level it seems to have clear value for describing what people mean by assigning causal status.

Causal reasoning goes beyond the criteria for considering candidate causes. It involves formulating arguments about the potential causes. The next section reviews some traditional assumptions about causal reasoning.

1.2 Claims about Causal Reasoning

We have identified five assumptions about causal reasoning that appear in many forms and we believe they are widely held. We also believe they are misleading. We do not claim to be the first to voice these objections; we are repeating them because the assumptions are still widely held.

1.2.1 Philosophy is the Basis for Understanding Causal Reasoning. Legions of philosophers have helped to illuminate the nature of causal reasoning. However, these illuminations generally center on the necessary conditions for valid or rational causal reasoning, with rationality set in terms of the standard of logic. In real-world settings, the evidence for causation is typically too ambiguous to permit valid (i.e., deductive) reasoning, so this is not a generally useful standard. Our goal is to describe how people such as military leaders and managers actually engage in causal reasoning. They are rarely in a position to satisfy the criteria for valid causal inferences and the problems they deal with do not fit neatly into manageable packages or fixed structures.

1.2.2 The Scientist is the Ideal for Causal Reasoning. Much of the literature in cognitive psychology and in the psychology of science focus on causal reasoning on the part of scientists, especially about physical causation (e.g., Gopnik & Schulz, 2007; Sloman, 2005). However, researchers, at least those who are engaged in the so-called 'hard' sciences, usually undertake investigations into determinate problems where there is a chance of making a discovery. The investigation into the cause of acquired immune deficiency syndrome (AIDS) led to the discovery that human immunodeficiency virus (HIV) causes AIDS. Watson and Crick (1953) figured out the structure of deoxyribonucleic (DNA) and explained how DNA could replicate. In contrast, military leaders, organizational managers, and researchers in the 'soft' sciences ponder indeterminate questions. Why did the American military situation in Iraq improve from 2005 to 2008? Why did Hillary Clinton lose the contest to become the Democratic candidate for president in 2008? Why did a certain sports team (name your favorite example) beat another in a championship game? There are no single or uniquely correct answers to such questions, and no amount of research would discover the "real" cause. A model of causal reasoning that fits the kind of reasoning that 'hard scientists' engage in does not fit causal reasoning in general.

Scientists are driven by curiosity and are always looking for deeper explanations and further mysteries, whereas managers have to stop at a certain point and make decisions. Feltovich et al. (2004) describe a "reductive tendency" to chop complex events into artificial stages, to treat simultaneous events as sequential, dynamic events as static, nonlinear processes as linear, to separate factors that are interacting with each other. Scientists are on the lookout for these tendencies, whereas managers, leaders, and other kinds of decision makers depend on the reduction to avoid some of the complexity that might otherwise be unleashed. Therefore, the scientist, working on deterministic problems and searching for deeper and deeper explanations, is an inappropriate model for naturalistic causal reasoning.

1.2.3 Causal Reasoning Means Finding the True Cause for an Effect/Event. As described above, when dealing with indeterminate causes there is no way to identify a single or "true" cause. Further, researchers such as Reason (1990) have shown that accidents do not have single causes, so the quest for some single "root" cause or a culminating cause is bound to be an oversimplification and a distortion. Nevertheless, in order to take action we often need to engage in such simplification.

1.2.4 Correlation Does not Imply Causality. But of course it does; it was designed to. Correlational studies are often taken to demonstrate causal relations. Correlation, as a suite of mathematical techniques, was invented precisely to enable the exploration of causal relations or potential ones. Correlation is a major cue to causality. Even in scientific investigations, correlation is required in order for causation to be proved. The source of confusion here is the term "implies" which can mean "suggests" or "requires." Correlation certainly suggests causality, but it does not require a conclusion of simple causality. Further, people do not mentally calculate correlations, but rather apprehend co-occurrences and covariations. Sharp observers use coincidence to speculate about causality. The coincidence of prevalence/absence of mosquitoes and presence/absence of yellow fever helped control and then understand the disease. It is true that correlation does not prove causation, and that additional factors may be operating and causing both the putative cause and the effect. But correlation definitely suggests causation. It often initiates a fruitful causal investigation.

1.2.5 In Causal Reasoning People Identify an Effect, Nominate Causes, and Select What They Believe is the Best One. This approach fits scientific investigations. It does not always fit medical investigations. After all, the original AIDS "effect" to be explained was why gay men were dying of infectious diseases. As the investigation continued, the perceived effect morphed to include intravenous drug users, then also people who had received blood transfusions, and other at-risk populations. Cases such as these show that the initial effect may be re-framed and re-cast during the investigation into its causes.

Summarizing the limitations of these claims we see, that in many natural settings involving human activity, the causes are often multiple, vague, and indeterminate. Frequently people never figure out actual or final causes. People sometimes stop their investigations at a fairly shallow level, demonstrating the reductive tendency. The effects we are trying to explain morph. Time lags between cause and effect are inevitable; Time lags create an additional layer of complication (Dörner, 1989), not simply because of the time but because of intervening events that cloud the picture. People still have to engage in causal reasoning under these conditions, but their reasoning will not follow the models of philosophers and scientists of leading to some single, final point where causal reasoning stops because the cause has been determined and the explanation of events is complete.

Causal reasoning can take different forms in natural settings and is rarely amenable to factor analysis of causal chains. Our research examined the diverse forms of causal reasoning in naturalistic settings.

2.0 CAUSAL REASONING RESEARCH STUDIES

Our strategy was to pursue two issues in parallel – to explore the nature of causal reasoning and to design and develop an EUG for fusion algorithms and simulations models. These strands interacted: the causal reasoning guided the design of the EUG and the evaluations of the EUG shaped our models of causal reasoning.

We will describe the causal reasoning strand first, and then the EUG strand. And we will present the findings after each method, rather than having one large Method section followed by a disconnected Findings section.

2.1 Causal Reasoning Study: Phase I

The *causal reasoning strand* was pursued in different ways in each of the three phases of the research project.

We began by integrating the literatures on causal reasoning by compiling a master list of all the published lists of types of causes and cause-effect relations. We also analyzed the functional purposes of causal reasoning (Table 1).

Table 1: Functional Purposes of Causal Reasoning

	Observative	Agentive
Prospective	Reasoning about what someone else thinks will happen.	... in events in which either agent has a causal power
Interventive	Deliberate experimental action to probe the cause-effect relation or test some theory	... in events in which either agent has a causal power
Inspective	Reasoning about what someone else thinks is happening.	... in events in which either agent has a causal power
Retrospective	Reasoning about what someone else thinks has happened.	... in events in which either agent has a causal power
Disruptive (e.g., garden pathing; deception)	Reasoning to influence someone else's reasoning, e.g., deception	... about events in which either agent has a causal power
Preventative	Reasoning to prevent someone else from engaging in causal reasoning	... about events in which either agent has a causal power
Corrective	Recognition that there is an explanatory gap. Reasoning about what went wrong in someone else's causal reasoning. Responsive gap-filling (Response to recognition of a Black Swan)	... about events in which either agent has a causal power
Protective	Reasoning to achieve a justification of rationalization of someone else's (or some organization's) actions, to provide a rationale (e.g., "cover your neck" and "scapegoating")	... about events in which the reasoner or the organization has a causal power

The data collection in the first phase was based on a textual analysis of published materials in domains such as political, military, economic, and social explanation for events.

Our first step has been to collect newspaper, magazine and book materials illustrating causal reasoning about natural events. We collected 74 stories from newspapers and news magazines with the goal of sampling varied venues of human activity including sports, politics, world events, and economics. The sub-prime mortgage crisis provided many explanations as the debacle unfolded. The 2007-2008 American football playoffs and Super Bowl offered different types of accounts. The Republican and Democratic primaries generated ample speculations about the reasons why different candidates succeeded and failed. The changing conditions in Iraq stimulated analyses of what went right and wrong.

In studying accounts that presented causal analyses and causal explanations, we identified the individual statements of causal attribution, we labeled the statements with identifiers, and we made notes that summarized each attribution. As we collected and analyzed more accounts, we began to see some convergence of themes. In some explanations the cause was seen as a single dramatic event that could have gone the other way (e.g., a basketball team lost a game because of a basket at the very end of a game); whereas, in others there was a critical event but it was not so dramatic, coming earlier in the event sequence. Both of these a theme of the single critical and reversible event, we found accounts that seemed to boil down to a single condition for how something could happen (e.g., HIV causes AIDS), but we also found stories in which the mechanism was complex, involving multiple causes in which the effects interacted with one another.

For instance, one story offered an explanation of the increasing cost of products made in China (the effect X to be explained). Some causes led directly to the effect. For example, China reduced and removed tax incentives for exporters of Chinese goods (A), which led to increased costs of exports (A→X). Product recalls and environmental crackdowns (B) also led to increased cost of products made in China (B-→X). Causes were also indirect. For example, an increase in oil costs (C) led to an increase in the cost of plastics (D), which led to an increase in the cost of Chinese products (C→D→X). Labor shortages and stricter labor rules (E) led to an increase in wages, which (F) led to an increase in the cost of Chinese products (E-→F-→X). This seemed to be a "swarm" of converging effects but it had some chains of effects.

We should also add some comments about coding the data. In an informal check on coding reliability we found that domain knowledge seems critical for reliably identifying the causes mentioned in a media account. However, once the causes are specific, domain knowledge does not seem necessary for coding the causes into the explanatory categories shown in Table 2.

As shown in Table 1, the pilot study focused on an analysis of 74 incidents. These often referenced more than one cause; we tallied 219 individual causes. Only two of the 39 sports incidents referenced 10 or more causes. In contrast, four of the 18 economics incidents included 10 or more causes. None of the political, military or miscellaneous incidents had even 10 causes. In each account, we identified the causes as the reasons offered by the writer to explain why the outcome occurred. (The column to the extreme right in Table 2 shows the number of cases that mentioned any contrary information – influences that worked in the opposite direction of the outcome. This category was not one of the causal reasoning themes.)

Table 2: Causal Reasoning Themes

Topic	# Cases	Events	Abstraction	Condition	List	Story	Counter
Sports	38	17	55	29	14	12	12
Economic	18	20	3	25	2	14	3
Politics	7	17	5	14	7	0	2
Military	3	4		4	2	0	1
Miscellaneous	8	18		8	1	3	0

From these data, we identified five causal explanation themes: Events, abstractions, conditions, lists, and stories.

2.1.1 The Theme of Events. These were mutable, that is, reversible events, actions or decisions, sometimes referred to as counterfactuals or close counterfactuals. For example, late in the last quarter of the 2008 Super Bowl between the New York Giants and the New England Patriots, Eli Manning, the Giants' quarterback, seemed almost sure to be sacked by the Patriots but somehow spun away and got off a pass that the receiver caught against his helmet. Most accounts of the game highlighted this miracle play because if Manning had been sacked the game would probably have ended with the Giants losing, and it was very easy to imagine the play failing. Events that are mutable are more convincing causes than routine events – in basketball, winning a game with a free throw in the first quarter less is likely to be cited as an explanation than winning with a 3-point shot. A last minute reversal, such as the winning 3-point shot, appears to be both necessary (the team would have lost if the shot missed) and sufficient (at that point, the game was completely decided by the shot).

As would be expected, the sports incidents included a large share of these kinds of reversal (counterfactual) explanations. In the economics category, the United States (U.S.) Federal Reserve decision to keep interest rates low in the period 2002-2004 has been identified as a cause of the housing boom, the housing bubble, and the subsequent recession.

2.1.2 The Abstraction Theme. This form of reasoning takes several causes, including counterfactuals, and synthesizes these into a single explanation. In basketball, a series of mistakes by the New York Knicks (a professional sports franchise) were synthesized to explain why the Knicks lost the game. Table 1 shows the Abstraction theme was more prevalent for sports than for economics.

2.1.3 The Condition Theme. This explanation cites a prior condition even before the to-be-explained event began. Thus, in sports, if a key player was so injured that he did not even play, we counted that as a Condition because it did not occur during the contest. Economics offers many examples of conditional explanations – a market force inexorably at work, such as the development and collapse of bubbles. Often, a conditional theme is used in a simplistic fashion. The economic recession is blamed on greed. The success of a sports team is attributed to better coaching, or the fact that they "wanted it more." Or consider the cause of World War I. The assassination at Sarajevo explains it as an event, whereas the rise of nationalism explains it as a condition – a feature of the situation. We are using this category to include lawful relationships and regularities, as well as characteristics of a situation such as an injury to a star player. We suspect this category of "condition" may evolve in the future.

Sometimes, these three types of explanations (an event, an abstraction, and a condition) were offered by themselves, but often they were bundled together. We identified several common ways for them to be bundled into a higher-level explanation: the abstraction, the list and the story. The category of Abstraction is sometimes offered by itself, with exemplars being implicit, but at other times the Abstraction was used as an additional way to bundle events in which all of the relevant factors and events are of the same kind. Most important, an abstraction is usually offered as a single answer to the question of what caused an event, in contrast to lists and stories.

2.1.4 The List Theme. This is merely a list of multiple reasons why something happened and converged. Lists are fairly common in sports – e.g., the reasons the Patriots lost the Super Bowl. For the sports category, 14 of the 38 accounts featured a list. Lists are less common in economics – an example would be an article listing the reasons why the Chinese economy should move into a higher rate of inflation. All of the articles on politics relied on a list – the reasons the political campaigns of John Edwards, Rudy Giuliani, Mitt Romney, or Hillary Clinton folded.

2.1.5 The Story Theme. This provides a deeper analysis to present a mechanism of how the different causes interacted. Sometimes the stories took the form of a chain. These kinds of chains were relatively rare in the sports incidents, and when they were used, the chains were very short. Chain-reaction stories seem more prevalent in economics. In general, economics analyses used the most complex story explanations. For example, one article described how the Federal Reserve worsened the sub-prime mortgage problem. An example of such a story is presented in Table 3. The story here describes different causes acting in parallel, but also interacting, to produce the conditions for an economic crisis in the U.S. The causal analysis tried to explain why the Federal Reserve made a critical mistake in 2002 when it continued to reduce rates.

Table 3: The Story Explanation for the U.S. Mortgage Crisis

• In January, 2001 the .com bubble was bursting.
• A recession was starting.
• After the 9/11 attacks there was a fear of deflation.
• Therefore, the Federal Reserve cut the rate by .5% outside the normal schedule for announcing rate changes, from 6.5% to 6%, followed by 12 more cuts through 2003, dropping the rate 5 points, eventually to 1%, the lowest rate since 1958. The Federal Reserve kept rates at this level for a year. Then the Federal Reserve increased rates by ¼% increments, to 5.25% in June 2006.
• However, by 2002 it was clear that rates should be staying neutral or going up, not down. o The Gross Domestic Product was lining up with capacity. o Inflation was low.
• In addition, the housing market was vibrant, even in 2001. o Housing does not always follow the law of supply and demand. When prices rise, that creates a demand in the form of a bubble as people expect prices to keep rising. o The rise in housing prices created an increase in house building, strengthening the economy.
• Mortgage rates stayed low even when the Federal Reserve under Bernanke, raised the rates. o The reason was Bernanke suggested the low rates were a global issue. Oil exporters and thriving Asian economies needed places to invest. o This added to the money supplies in the U.S. and kept rates low.
• Lending standards were reduced, which is typical in a bubble/craze.
All of these credit-cheapening forces helped the sub-prime borrowers enter the equation, as looser practices and pressures enticed less-qualified investors.

Here is another example of a story, from the Miscellaneous Topics group of incidents. The effect was the death by asphyxiation of a fireground commander in New York. How did it happen? This story is presented in Table 4. Each of the elements of this story is a cause – each was reversible, each led to the subsequent events/effects.

Table 4: The Story for the Fatal Blaze Incident

• A woman in an apartment was giving her children baths and wanted the apartment to be warm so they wouldn't catch colds. The apartment was already adequately heated. She increased the temperature by turning on the gas stove.
• Her young son, waiting for his turn, started playing with a paper wrapper from a toy. He waved it over the flames and it caught fire.
• He became frightened and tried to hide it behind a sofa.
• The sofa caught on fire and the flames spread.
• The mother came out of the bathroom, saw the flames, gathered up her children and fled the apartment.
• On her way out, she dislodged the rug by the front door.
• The rug got stuck in the self-closing door, preventing it from closing.
• In her rush she didn't check the door.
• Because the door didn't close, the fire and smoke were not contained to the apartment. They spread into the hall.
• Shattered windows created winds that fanned the flames.
• The firefighters arrived and were thwarted by low visibility because the hall was filled with smoke.
• Their progress was so slow that they began to run low on oxygen from their tanks.
• Accordingly, they had to withdraw.
• The unit leader failed to withdraw. o Perhaps he was still searching for residents. o Perhaps he wanted to be sure that everyone in his crew had left. o Perhaps he became disoriented.
• He ran out of air and died.

Here we see a set of causes related to the spread of the fire into the hall, and another set about the failure of the lieutenant to withdraw in time. There is no single event or simple sequential chain. Moreover, confidence in a causal chain or interaction depends on the plausibility of each transition. In this regard, it is not clear how to treat violations of expectancies. If a transition is highly plausible then it should add to confidence but also diminish the information value of the account. Transitions that violate immediate expectancies but seem to be well-justified may increase confidence in the account.

2.1.6 The "Counter" Theme. The "counter" theme simply tabulated the number of incidents in which an opposing cause was mentioned – e.g., a good play by someone on the losing side of a sports contest. These counter-explanations emerged fairly often in sports, but were rare in the other categories.

The greater complexity of the explanations in economics, as compared to sports, should be caveated somewhat. The economics explanations created an impression of inevitability, a sense that this is how the dominos were fated to fall. The mortgage crisis example is an anomaly that suggests the Federal Reserve should have acted differently and might have altered fate. Most economics explanations fail to include any countervailing forces or opportunities for events to unfold differently. They are perceived to be strongly determined. In contrast, many of the sports accounts note countervailing causes. A few of the Super Bowl accounts note the Giants were lucky with their miracle play which changed the outcome of the game. Of the 38 sports incidents, 12 cited some sort of countervailing force. Only 3 of the 18 economics incidents did so. Sports accounts seem to be more sensitive to factors such as luck, and sometimes offer a counterfactual perspective that is usually missing from economics.

Fugelsang, Thompson, and Dunbar (2006) have referred to the "list" theme as "multicausal," and, within story explanations, distinguished domino chains as "linear" and more complex stories as "interactive." However, their use of "multicausal" involves only necessary conditions (e.g., for a flower to bloom it must receive sunlight, fertilizer, warm temperature, and moisture), whereas our analysis of lists includes factors whose influence is not fully determined, such as the reasons that Hillary Clinton's presidential campaign failed.

We formulated a Butterfly model of causal reasoning – so named because its shape resembled a butterfly (Figure 1).

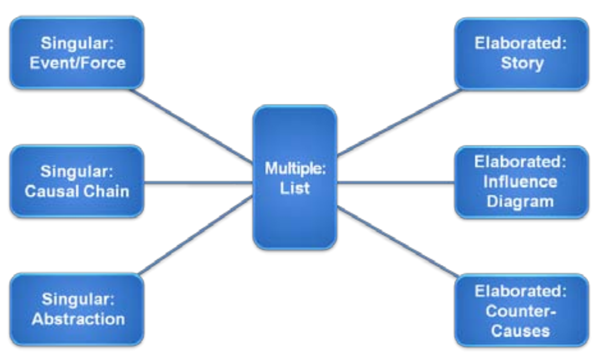


Figure 1: Butterfly Model of Causal Reasoning

2.2 Causal Reasoning Study: Phase II

Hoffman followed up on this study in Phase II of the project. This study, conducted at the Institute of Human Machine Cognition, used a set of media accounts of events and examined the ability of participants to identify the causes and to categorize these.

2.2.1 Procedure. Five brief news articles were selected, representing the domains of sports, world events, history, and economics. Each article was one page or less, and they spanned several different domains.

1. Economics. "How Did Madoff Get Away With it?" *The Week*, January 2009.
2. History. "Supply and Demand: The Industrial Revolution Explained and Why Britain Got There First." *The Economist*, Fall 2009.
3. Sports. "Patriots Not So Perfect After All." *Pensacola News Journal*, January 2008.
4. Business. "Inside Every Chief Exec There is a Soviet Planner." *The Observer*, 2009.
5. World Events. "Can Obama Tame North Korea?" *The Week*, November 2008.

Sentences, clauses, or phrases in each article that represented causal attributions were identified and underlined by the Researcher, using the conceptual terminology of the causal themes: abstraction, enabling conditions, counter-cause, reversible event or decision, list, or chain/domino. Examples are:

- *"The SEC got multiple warnings but failed to uncover Madoff's activities."* (Counter-cause)
- *"The Industrial revolution began in Britain because it was profitable there and it fit a demand."* (Abstraction)
- *"Steam engines had been designed to pump water, watchmakers provided high-quality gears…."* (Chain/List)
- *"Why did Britain have high wages and cheap energy in the first place?* (Abstraction)
- *The Patriots won the Super Bowl because their receivers kept catching Manning's last-ditch passes."* (Abstraction)
- *"Relations with North Korea worsened because Bush tried to bully Kim."* (Reversible)

The labeling was an emergent process, as the preparation of materials of this study also fed into our continuing development and refinement of the themes. For instance, in our initial appraisal of the process of causal reasoning we discussed the phenomenon of "onioning" in which the reasoner asks about the cause of a cause, but we had not included that in our brief list of themes because it seems to rare, and could be counted as a type of Reversible or a type of Abstraction. Also, we had included the Chain as one of the themes but the examples we found in articles were typically lists of factors rather than chains of cause-effect relations.

The experiment was run as a self-administered booklet, presented to 11 psychology majors at the University of West Florida. Each article was printed on a separate page, and written at the top was the effect to be explained (e.g., *"What causes the U.S.-North Korea hostility?"*). Instructions were to read the article and then go back and underline what were seen as causal attributions. Next, each underlined attribution was to be rationalized by answering the questions:

"How does this explain the cause?"
"Is the story-teller's explanation simple or complex?"
"What is your level of prior knowledge and interest in this topic?"

After reading a relatively simple worked-out example, the Participants worked through their booklets, time-stamping the turn of each page. It generally took Participants about an hour to an hour-and-a-half to finish the booklets.

2.2.2 Results. Eleven participants completed the experiment. Table 5 shows the results with respect to the articles: The statements that Participants underlined were also statements the Researcher had underlined and categorized. These might be thought of as "hits."

Table 5: Results on Participants' Underlining of Causal Attributions that were also Noted as Causal Attributions by the Researcher

Story	Total # of Attributions Seen	Total # of Opportunities to See	Mean (over Participants) Proportion of Attributions Seen
Industrial Revolution	29	165	0.24
Madoff	37	77	0.51
North Korea	41	154	0.36
Patriots Win	28	99	0.18
Planners	41	231	0.18

This shows that for all of the articles, the Researcher identified more causal attributions than all of the Participants.

Table 6 shows the results with respect to the Participants, the average proportion of statements each Participant underlined that were also statements the Researcher had underlined and categorized.

Table 6: Average Proportion of Hits for Each Participant, Averaged over Attribution Types within Articles

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
Sum	42	28	19	13	20	12	19	33	17	21	21
Mean	0.43	0.29	0.25	0.20	0.20	0.10	0.30	0.30	0.18	0.34	0.24

Participants tended to notice this only about one-third to one-fifth of the number of potential ("seeable") causal attributions.

Table 7 shows the results with respect to the types of causal attributions, the "Seeability Proportions." Seeability was the total number of times that participants underlined an attribution relative to the total number of times the attribution was presented to participants. Thus, if an article had two Reversibles, and there were 10 participants, the total number of opportunities for seeing would be 20.

Table 7: Seeability Results

Causal Theme	Occurrences (Number of attributions in all the articles)	Opportunities (Participants x Occurrences)	Average (over participants) of the Proportion Seen
Enabling Condition	23	253	0.15
Reversible	30	330	0.30
Counter-Cause	7	77	0.21
Abstraction	30	30	0.37
List/Chain	2	22	0.32

All of the Participants "saw" Reversibles and Abstractions. (We say "saw" because the Participants did not rely on our conceptual terminology.) All but two of the 11 Participants "saw" Enabling Conditions.

Only 2 of the 11 Participants underlined Counter-causes. There were only seven Counter-causes in the articles, and hence 77 opportunities. Six of the participants did not underline the List/Chains. There were only two in the articles, but going into the study, we thought these were salient and obvious. For instance, in the article on the causes of the Industrial Revolution there was the Chain: "*The Black Death raised the price of labor and boosted trade, for English sheep grew longer fleeces, and local cloth improved. As Britain traded more... other cities expanded.*" We expected all Participants to underline such obvious chains, but only 6 of the 11 did, and even then they did not uniformly underline all of the elements within in the chain or list.

One possible outcome was that Participants could underline statements, regarding them as causal attribute that the Researcher had not highlighted and labeled in the master scoring key. Out of 245 underlinings by all the Participants, there were only 19 such additions, which contrasts with the total of 92 attributions identified by the Researcher.

Of the 19 additions, 8 were from just one of the articles. Most of the additions were either restatements of the effect that was to-be-explained or were causal attributions only by inference. An example of an effect restatement is "*Madoff stole money,*" which was regarded by one Participant as a causal explanation. Arguably, it could be thought of as an enabling condition and, thus, we would consider it a causal attribution that the Researcher missed. An example of attribution by inference is the statement "*We played them like they were any other team,*" which was interpreted by one Participant to mean that "*They were not intimidated.*" While this could be regarded as an Enabling Condition, we did not count it as a "Researcher miss" because the attribution was by inference, that is, it was not explicit in a statement in the article.

Combined with the results on "hits," it seems clear that the Researcher not only identified many more attributions than the Participants but identified most of the attributions that people might find. Most people can "see" Reversibles, Abstractions, and Enabling Conditions. The Participants explanatory comments show that they apprehend the justification (or explanatory value) of an attributions in ways that accord with our conceptual terminology. Examples are presented in Table 8.

Table 8: Examples of Attributions and the Participants Explanations

P	Article	Underlined Statement	Researcher's Categorization	Participant's Explanation
1	**Industrial Revolution**	*"... the conditions were not sufficient or exclusive to Britain..."*	Counter-cause	*"These explain why the author thinks that previously considered factors are insufficient."*
1	**Industrial Revolution**	*"The Black Death raised the price of labor and boosted trade, for English sheep grew longer fleeces, and local cloth improved. As Britain traded more... other cities expanded...."*	Chain/List	*"Long series of C-E statements to explain why wages were high and energy cheap."*
4	**North Korea**	*"North Korea fears that it has become a low priority for an incoming Obama administration that will have to deal with other issues."*	Abstraction	*"NK knows that the Obama administration has too many problems to tackle and that they are towards the bottom of that list. So they are creating hostility with the USA for attention."*
5	**Patriots Win**	*"When Eli Manning put up last-ditch pass after last-ditch pass, and his receivers kept catching them…"*	Abstraction	*"Explains how Giants QB made many last ditch long passes, causing Giants to gain momentum."*
6	**Planners**	*"If anything, overhead costs are increasing as work breeds more work. In less effective organizations, of course, hidden indirect costs are much higher. "*	Enabling Condition (Spiral)	*"Overhead increases as work breeds more work—hidden indirect costs are much higher."*
7	**Industrial Revolution**	*"The industrial revolution occurred in Britain in the 18th and early 19th centuries for one overwhelming reason, he argues: it was profitable there and then. It met a demand."*	Reversible	"Profit caused the IR to occur in Britain."

Many of the Participants' justification statements were explanatory elaborations, a good example being row three in Table 8. Also frequent were justifications that essentially restated article statements. A good example is the last row in Table 8.

We noticed some interesting individual differences, constituting one of the findings that motivate further investigation. We pointed out in Table 8 that participant performance ("hit" rate) ranged from about 50% down to about 10%. There are many explanations for this, and for the discrepancy of 245 vs. 92 comparing the Participants and the Researcher. Many of the possibilities will involve individual differences. That it is possible using this experimental paradigm to observe a range of performance means this paradigm could be used to explore the individual difference factors. For example, one Participant did not see any causal attributions in one of the articles (the sports article) and yet this participant saw all of the types of themes in other articles. This same participant said that a causal explanation in one of the articles was "too complex for most people" and yet also said that there were explanatory gaps.

We are in the process of collecting additional data with the participation of experienced intelligence analysts.

2.2.3 Textual Analysis. In Phase II we also continued our textual analysis but in a more focused manner, by conducting more in-depth examinations of issues such as the reasons why Xerox failed to commercialize the personal computer breakthroughs it achieved; the causal reasoning involved in understanding diseases such as AIDS, yellow fever, and cholera, and the causal reasoning in understanding a friendly fire incident in which two U.S. Air Force F-15s shot down two U.S. Army Black Hawk helicopters in northern Iraq in 1994.

2.2.4 Cognitive Task Analysis. During Phase II our research on causal reasoning expanded from textual analysis to Cognitive Task Analysis (Crandall, Klein, & Hoffman, 2006). Our review of textual materials illustrated how outsiders, with ample amounts of time, considered the causal dynamics for events, but we needed to understand how decision makers engaged in causal reasoning under conditions of time pressure, uncertainty, and so forth. Therefore, we conducted 10 2-hour interviews, using the Critical Decision Method, with specialists in logistics, intelligence, and command and control. Each interview was conducted by a team of two to three researchers, and one subject-matter expert (SME) at a time. The interviews were primarily conducted at the offices of the Klein Associates Division of Applied Research Associates in Fairborn, OH, but some were conducted at Wright-Patterson Air Force Base and at the Institute for Human Machine Cognition in Pensacola, FL.

The Cognitive Task Analysis interviews helped us deepen this model and enabled us to explore additional facets of causal reasoning; in particular, our interviews with the logistics specialists. We expected them to grapple with the question of when to stop probing for more information about causal relationships, and we were surprised to find that we were off by 180 degrees. Rather than struggling with a stopping rule, they seemed to lack a starting rule. That is, they usually did not initiate any investigation into causal dynamics even in cases where such an investigation would be useful and enlightening. In one example, an Air Force logistics specialist was responsible for maintaining safety at an AF base; which included maintenance of the fire trucks. He had two larger tankers for spraying water on a fire. One was in poor shape and the other was in good mechanical shape. The one in good shape was scheduled for routine depot maintenance. He and his staff agreed the other one should go in for the routine maintenance, but when they requested this shift, the depot staff turned it down. End of story. We asked him why they turned him down, and he admitted that he did not know. It never occurred to him to try to find out. We speculate that this lack of curiosity may stem from inexperience – he did not see any value of finding out the reason why his request was rejected. In the interview we noted that if the depot did not want to extend them, perhaps his colonel might have been able to intervene, whereas if the rationale was a concern over legal repercussions in case the fire truck originally scheduled for maintenance was to run into mechanical difficulties then it might be harder to convince them otherwise. He agreed that at this point in his career, with a better appreciation for how to get things done, he might have pushed further. At the time, he was insensitive to any value of trying to learn why the request was denied. In virtually all of the interviews with logistics specialists we encountered the same phenomenon – a disinclination to pursue an inquiry into the causes for decisions.

The logistics interviews also surfaced an issue we had not encountered in Phase I – the need for causal inquiry into the reasons for the decisions and actions of other people. Our limited sample of interviews suggests that the reason for the actions and decisions of others may stem from: lack of information, competing priorities, organizational constraints, and/or lack of motivation.

Because of these findings, we engaged in a dialog with Litman (2008), who has developed a variety of curiosity scales, so that we could better understand the role of curiosity in causal inquiry. We included one of Litman's curiosity scales in a study performed in Phase III.

During Phase II we expanded our model of causal reasoning (Figure 2). We specified a set of single cause formats: an event, a decision, a force, and an abstraction. We also specified alternative multiple connected formats: a chain, a diagram, and a more complex diagram with counter-causes. We also identified various factors that should influence people's preferences for one format over another: individual differences such as need for closure (NFC) and curiosity need for action, as well as perceived audience sophistication.

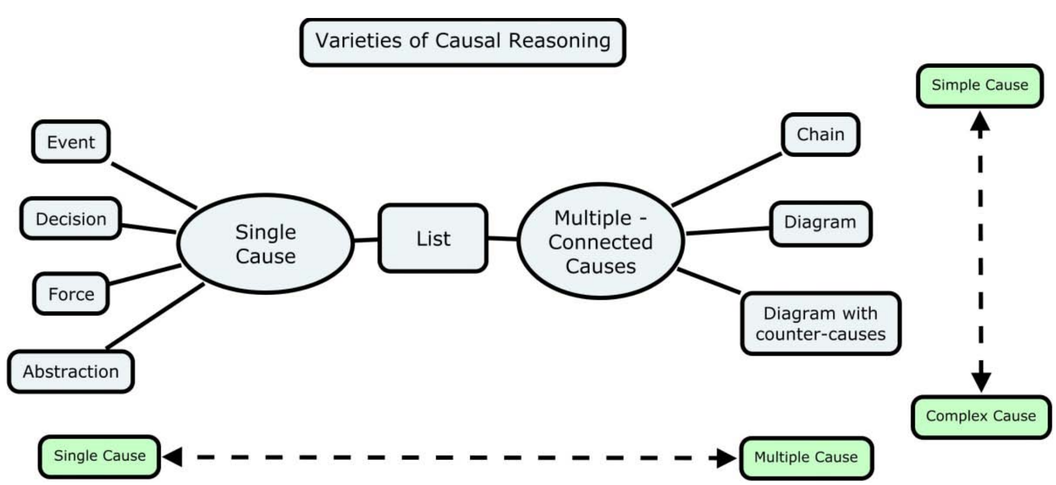


Figure 2: Butterfly Plus Model of Causal Reasoning

2.3 Causal Reasoning Study: Phase III

Phase III of the research on causal reasoning shifted to the collection of quantitative data from controlled laboratory conditions using college students. By Phase III we developed a sufficiently detailed model of naturalistic causal reasoning to warrant hypothesis generation and testing.

In Phase III we conducted two studies of preferences, to see if we could manipulate preferences in accordance with the factors shown in Figures 3 and 4.

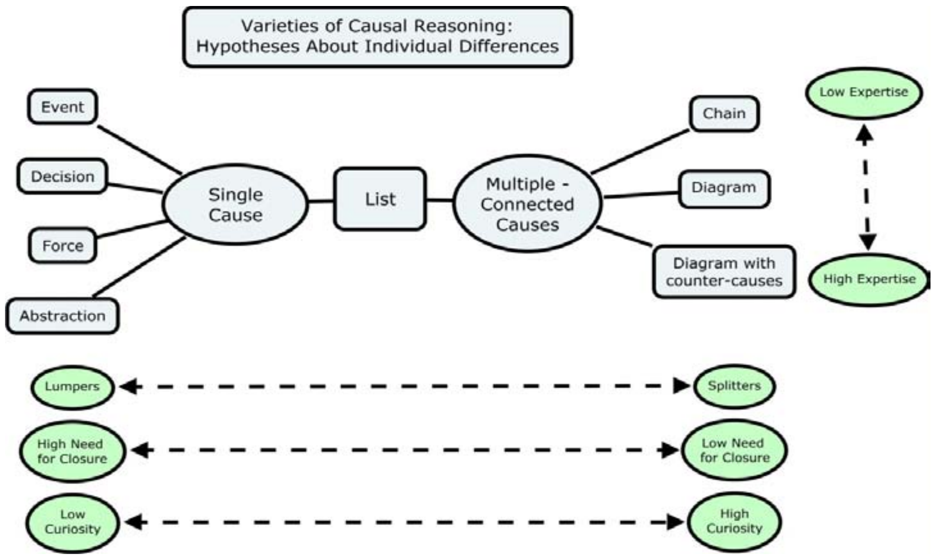


Figure 3: Butterfly Plus with Individual Differences

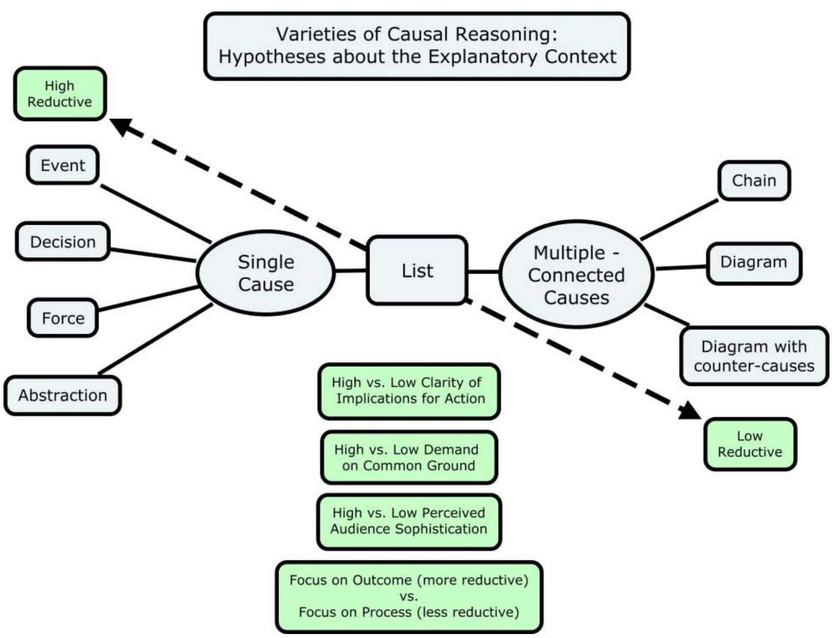


Figure 4: Preference Factors for the Butterfly Model

Wright State University in Fairborn, OH conducted a Phase III study. This study sought to manipulate preferences for different forms of causal explanations. If we knew how people engaged in causal reasoning, we should be able to manipulate their preferences for different forms of causal reasoning. We presented two different scenarios, one involving alternative explanations for the U.S. economic collapse in 2007-2008, the other involving alternative forms of explanations for the relative success of U.S. military forces in Iraq starting in 2007 (see Appendix A for scenarios).

We asked participants to rate their preference for different forms of explanations depending on whether they wanted to explain the events to someone who was young and inexperienced, or whether they wanted to understand what happened for themselves, or whether they wanted to be able to better handle such events in the future, or whether they wanted to provide advice to personnel who were engaged in making decisions about these kinds of issues (see Appendix B for list of explanations for both scenarios). We also collected data on two individual difference dimensions—curiosity and need for cognition using The Curiosity Scale (Litman, 2008) and the NFC Scale (Webster & Kruglanski, 1994).

2.3.1 Participants. The participants were primarily Wright State University undergraduates enrolled in an Introduction to Psychology class (N = 62). They participated in this study in return for credits towards completion of their class. A sub-set of Wright State University students was drawn from the Wright State University Reserve Officers' Training Corps (ROTC) program (N = 20). A third sample involved graduate students enrolled in the Master of Business Administration (MBA) program in the School of Business (N = 20). A fourth sample consisted of undergraduate students at Sunway College in Malaysia (N = 42). These were Chinese students and were included in order to determine if they preferred causal explanations that were more or less complex than the U.S. students. All of the data were collected in groups, with the participants individually filling out booklets asking them to rate preferences and to complete the personality inventories.

2.3.2 Materials. The experimental materials consisted of a booklet requiring written responses within three major sections. In the first section of the booklet we asked the participants to provide standard demographic information, i.e., age, education level, gender, major, knowledgeability about each of the events, and ethnic background.

The second section of the booklet prompted participants to provide information about their causal reasoning preferences. Within this section the participants were asked to evaluate a number of different, plausible explanations for why two high profile public events occurred. These events included the collapse of the U.S. economy in 2007-2008 and the recent success of the U.S.-led coalition in Iraq (2007-2008). The participants were instructed that no special knowledge about these events would be required. They were informed that they would not be asked to generate explanations for these events; instead they would be shown a series of alternative explanations and would then be asked to indicate which ones appealed to them the most. In the last section of the booklet, the participants were asked to complete the Curiosity Scale and the NFC Scale. It took the participants 30-45 minutes to complete the booklet.

2.3.3 Results and Discussion. We found that audience sophistication influenced explanatory preferences. As predicted, participants selected simple explanations 10.6 times more often than they did complex explanations when the audience was less sophisticated, e.g. when explaining the events to a nephew (Figure 5).

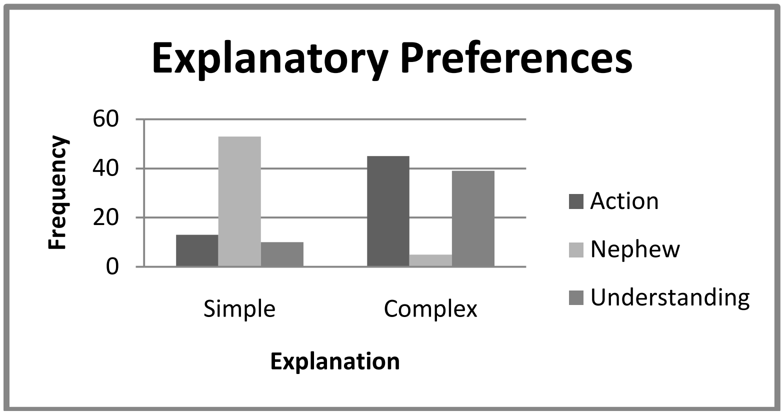


Figure 5: Explanatory Preferences, General Population (N = 40).
Simple = Explanations 1-4, Complex = Explanations 6-7 for Iraq and 6-8 for Economy

In a second version of this study, instead of examining explanatory preferences for less sophisticated audiences, we looked at preferences in the context of understanding and of action. We found the *understanding* context resulted in a slight preference for complex explanations relative to the *action* context. Complex explanations were selected 3.9 times more than simple ones for the understanding context, and 3.5 times more often than the action context. The relative frequency that complex explanations were selected compared to simple explanations did not differ by understanding or action context, $X^2(1, N = 309) = .84$, $p > .5$.

In this study, in addition to collecting data from the general population (N = 20), we also collected data on students enrolled in the University's ROTC program (N = 20), students enrolled in the University's MBA program (N = 20), and Chinese undergraduate students enrolled at a university in Malaysia (N = 43). As predicted, we found that Malaysian students showed a preference for more complex formats than American students did. The Chinese/Malaysian students (N = 43) selected complex explanations 8.2 times more often than simple explanations, whereas the American students (N = 60) selected complex explanations 2.8 times more often than simple ones. The U.S. and Malaysian populations differed with respect to the relative frequency with which the selected complex explanations compared to simple explanations, $X^2(1, N = 141) = 11.09$, $p < .001$ (Figure 6).

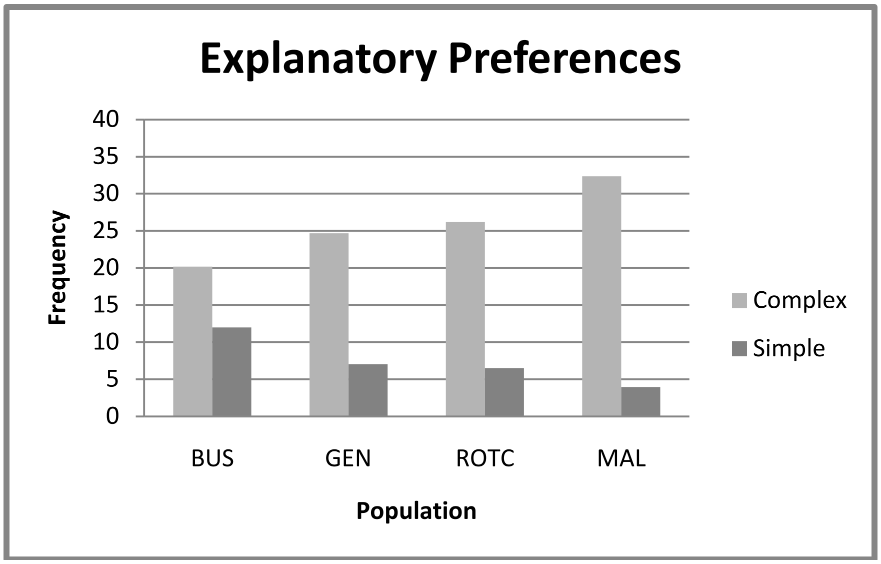


Figure 6: Explanatory Preferences across the Four Populations, across Recipients, Excluding Data for Option 5

There was no relationship between the individual difference measure of curiosity and explanatory preferences. The individual difference measure of NFC had a small effect, but in the opposite direction than we predicted – a higher need for cognition was linked to a greater preference for single-cause explanations. The need for cognition scores were 167 for single-cause explanations versus 162 for the complex explanations. This relationship was not statistically significant, however, $X^2(1, N = 143) = .37$, $p > .5$.

Our analyses divided the forms of explanation into three categories: "simple" (i.e., only a single cause), which was chosen in 11% of the cases, list, which was chosen in 43% of the cases, and "complex" (a chain or an influence diagram), which was chosen in 46% of the cases. There were four single cause explanations grouped under "simple." They identified an action, event, decision, or force. There were three forms of complex explanation for the economic scenario, a domino/chain, a simplified causal map, and a more comprehensive causal map. There were two forms of complex explanation for the Iraq scenario, a domino/chain and a causal map. Thus, Figure 7 compares just a single "list" against a basket of four simple and two to three complex causal accounts.

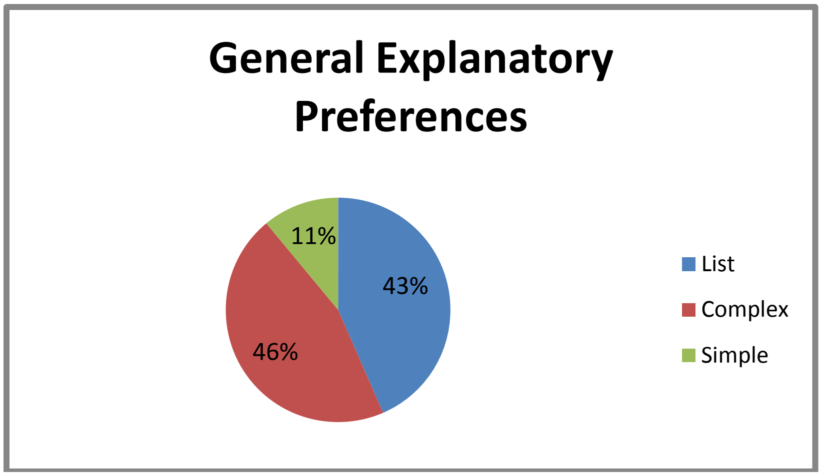


Figure 7: General Explanatory Preferences, across Populations, Recipients and Scenarios

Our analyses concentrated on the simple vs. complex comparison. The explanation that presented a list of possible factors or reasons seemed intermediate between simple and complex, and so we treated this option separately. In general, this "list" category was the most popular, more popular than any of the single forms of simple or of complex explanation. We did not find any systematic variation in preference for the list form of explanation across the two scenarios. The Business students did appear to prefer the list format over the simple or complex formats more consistently than did participants from the other populations (Figure 8). These data reflect first-choice preferences for two conditions: either an explanation that the respondent would prefer in order to understand the situation ('self') or an explanation that the respondent would prefer in order to take action ('invest' in the economic scenario, 'plan' in the Iraq scenario),

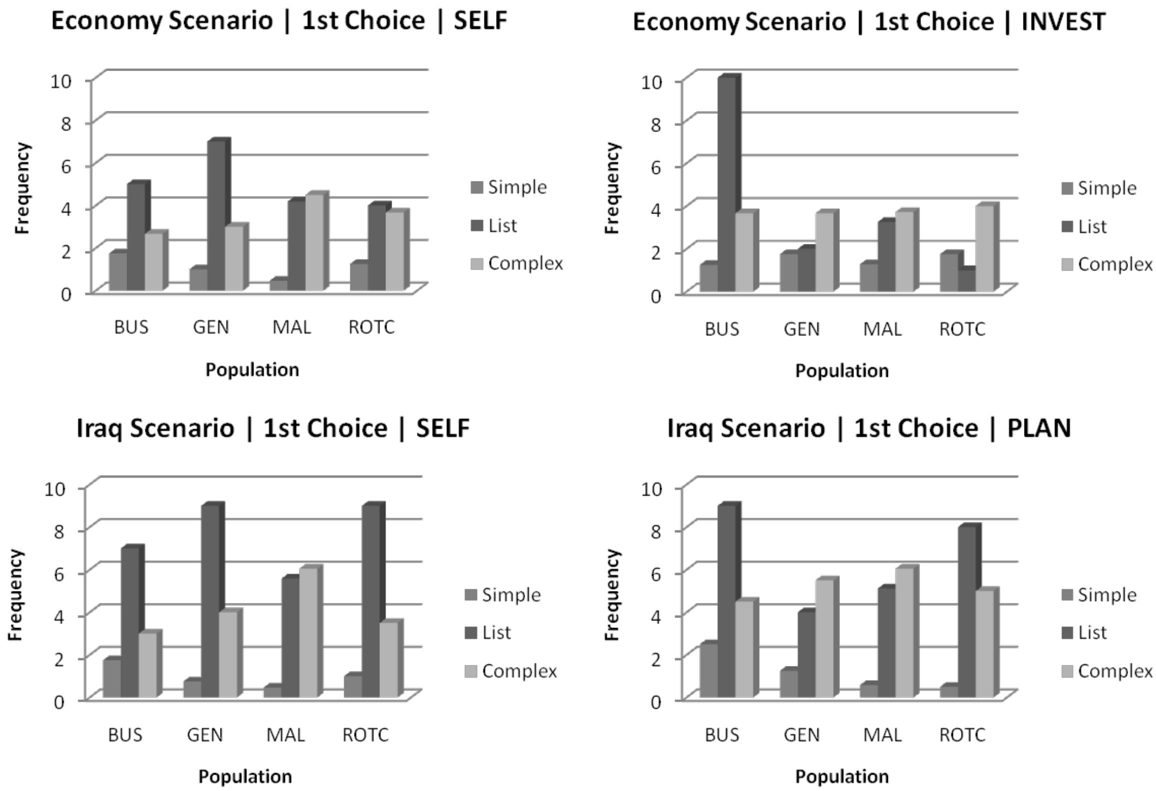


Figure 8: Explanation Types (Simple, Complex, or List Format) across the Four Populations. Dependent measure reflects frequency with which the explanation format is selected as the first choice

We also examined the effect of perceived familiarity with the topic (Table 9). Here, we were particularly interested in comparing the ROTC students with the MBA students because we expected that these students would have more knowledge of the surge in Iraq and the collapse of the U.S. economy, respectively, than students from the general population. Interestingly, the ROTC students reported being less familiar with events related to the surge in Iraq, than they were with events related to the collapse of the U.S. economy. Similarly, the business students reported higher familiarity with events related to the surge in Iraq than they did for the collapse of the U.S. economy. These findings are the reverse of what we expected. One possible explanation is that the more one knows, the more one appreciates one’s ignorance.

Table 9: Reported Familiarity with Experimental Events across the Four Populations

	EconFam	IraqFam
Business	2.60	2.90
Military	2.65	2.10
General Population	3.15	3.40
Malaysia	2.38	2.19

3.0 EXPERIENTIAL USER GUIDE

Our goal in the second strand of research was to develop a new concept for supporting users of intelligent software tools and other complex technology systems: the EUG. We will first describe some of the rationale the EUG, including the types of systems for which it is it most useful for. Then, we will report on some of generic components that make up an EUG. Finally, we will present highlights of an EUG we are developing for a Bayesian reasoning and modeling system called JCAT (Java Causal Analysis Toolkit).

3.1 Background and Rationale

There are many ways software developers document their tools to provide help to users. In the early days of personal computers, most software came with one or more heavy bound manuals which documented all of the features, shortcut keys, screens, and uses for the tool. As desktop computers have increased in power, these have been supplanted by help systems embedded in the software itself. These offer context-sensitive help, search, indexing, and tighter integration with the software tool, and often cost less to produce. Many embedded help systems document how particular buttons or functions of a software tool work, but leave out critical training about how to use the tool to solve typical and complex problems that users face. Sometimes, more advanced books and training is available for popular software titles, often developed by third party vendors. These more advanced training and user guides use a broad range of metaphors and strategies, many of which we examined to help develop the EUG concept.

To understand the properties an EUG should have, we evaluated approximately 30 types of user guides, documentation systems, and tutorial approaches used to help novice and advanced users of a very broad set of systems. Our goal was to identify aspects of user guides that are particularly successful in their domain, in order to inform the design of our EUG. Table 10 summarizes some of the highlights and lessons learned from this evaluation.

Table 10: Types of User Support Systems Evaluated to Identify the Properties of an EUG

System	Examples	Unique Features	Comments
"How to", "Hack" and "Maker" websites	Instructables.com, makezine.com, hackzine.com	Project-based, step-by-step how-to's and tutorials.	Highly experiential, focused on one-time projects.
O'Reilly Hacks book series	Google Hacks	Comprehensive set of advanced tips and tricks about specific tool/software/website.	Provide a comprehensive set of tips and tricks for a single tool or platform.
Gaming: In-game tutorials	Second Life's "Orientation Island".	Experiential set of tasks to understand system functionalities, tightly-integrated with automated mentoring.	In-game tutorials blur line between learning and playing, making them an attractive experiential guide.
Gaming: User-generated strategy guides	heavengames.com	Typically exists outside the gaming system; include screen-shots, screencasts, or text-based descriptions. Often task-based,	Does not require tight integration with software; deals with learning strategies to succeed in a

		challenge-based, or level-based.	complex environment.
Gaming: Walkthroughs	heavengames.com	Detailed shortest-path set of steps for accomplishing goal.	Can provide just-in-time support for tough problems.
Miniature Golf	The Joy of Cooking; Mastering the Art of French Cooking	Step-by-step ladder of challenges requiring skills of increasing complexity.	Natural sequences of lessons in EUG should be supported, but not required
Recipe Books		Basic guides for using a tool (the kitchen) to perform a task (create food).	Cookbooks range from targeted to general; are sometimes experiential.
Writer's Guides		Tells how to use a tool (word processor) to solve the problem (take plot ideas and create believable narrative).	By analogy, a writer's guide is to a word processor manual what the EUG is to a typical user's guide.
Introductory-level instruction books	Chess for Dummies, Complete Idiot's Guide to string theory, Excel for starters: The missing manual	Covers basics in a friendly way; does not alienate learner.	Typically targets highly-used consumer software with focus on most basic instruction.
Computer Help systems	Clippy, MSDN help	Embedded, context-sensitive, and can be comprehensive and include examples.	Implementations usually fail to live up to their potential, focusing on features rather than accomplishing tasks.
Automated Tutoring Systems	Circsim tutor (Evens & Michael, 2006)	Detailed analysis of whether learner understands complex set of causal relations.	Simulates some of the role of a one-on-one mentor.

One lesson we learned is that a user guide need not be tightly integrated into the software tool it supports in order to be successful. Many of the systems we looked at were developed by third-party sources, and often generated by users instead of the software developers or by technical writers. For example, on-line gaming communities frequently have large numbers of user-generated documents containing tips, strategy guides, walkthroughs, cheats, etc. These gamers typically cannot tightly integrate their help with the gaming software, and even in cases where this is possible (e.g., for games that allow users to generate their own content, levels, or scenarios), the skill and effort required to do so can be substantial, which also presents an obstacle that makes lower-tech alternatives more attractive.

Another source of user-generated help comes from "how to" websites, which offer tutorials for do-it-yourself projects (examples include *How to reuse a Walkman shell to hold an iPod*; *How to build an apartment-friendly bike rack*; *How to make a Valentine's day sock monkey*). These typically have step-by-step instructions and include photographs or video of the author doing the project. These "how to" websites also succeed by being low-tech, because highly-interactive content and video cannot be printed onto paper and used as a reference. These tutorials are also project-based, and so they offer an engaging experiential lesson in using the tools needed to accomplish the project.

Another lesson we learned is about the scope of different help systems. Some provide a comprehensive guide for advanced use of a tool or system, while others are targeted to a particular use, or a particular user set. For example, some cookbooks (e.g., *The Joy of Cooking*) aim to provide a comprehensive set of recipes, so that there are several options for almost any type of food the cook has available. Others target specific types of recipes (Julia Child's *Mastering the Art of French Cooking*) or specific types of users (*The Starving Student's Cookbook; The Everything Kids' Cookbook).* Similarly, user guides often target novices (e.g., *Chess for Dummies* and similar books in the *For Dummies* and the *Complete Idiot's Guide* series), providing detailed information to allow the user to gain enough facility to discover advanced uses on their own. Our belief is that an EUG should not try to be a comprehensive guide about the tool; rather it should focus on the places where there are genuine cognitive challenges for tool use.

Other help systems provide a learner-based organization, focusing on the sequence of tasks needed for achieving proper understanding. Tognazzini's (1998) example of how a miniature golf course provides its own user guide is an excellent metaphor: early holes exercise simple skills, which then get combined in increasingly complex ways. Similarly, automated tutoring systems such as the CircSim tutor (Evens et al., 2006) which trains medical students about the causal relationships between different aspects of the circulatory system, provide more adaptive methods to deliver the right lessons to the learner at the right time, while simulating some of the role that one-on-one tutors can provide. From these we learned the importance of identifying a sequence that the user should ideally use the EUG.

Although experiential training can be an effective way to support the learning of any complex task (e.g., learning to grow a lawn, or to drive a car, or to use a spreadsheet), there are some bounds on the types of tasks and systems we will focus on. Our focus is on intelligent software tools that pose real cognitive challenges to users. These are systems that perform non-trivial (Evens et al., 2006)reasoning, so that part of developing expertise with the system is understanding this reasoning. Potential users who do not understand this reasoning can become demotivated and can begin to distrust the results of the system or their own ability to use the system. This can lead potential users to reject a system that might actually be useful to them. We have identified four major aspects of systems that may pose non-trivial cognitive challenges to users: representation, data, computation, and output. These are defined and described in greater detail Table 11.

Table 11: Four Primary Functions of Intelligent Software Tools

Function	Definition	Examples
Representation and Modeling	How the system represents information, or how the user must generate a model of the world	Networks, vector spaces, trees, flowcharts, agents, rules.
Data Handling and Data Generation	How input for the system is generated.	Probabilities in Bayesian models; text corpora, sensor reliabilities.
Understanding Computation and Algorithms	How input are transformed using computational or mathematical processes.	Singular-value decomposition, Bayes rule, machine learning algorithms, fusion algorithms, optimization routines.
Output, Display, and Visualization	How the results are presented to the user to allow inference.	Visualization methods, tables for comparison; numerical scales, categorical classification (red-yellow-green).

Different tools and systems rely on different subsets of these functions. For example, a modeling tool will focus mainly on representation, and the primary cognitive challenge for the user is understanding whether the system's representation is similar to or different from his or her own conception, (and to some extent, whether it is similar to or different from how such a system would work in reality). Data handling and data generation differ across systems as well. For example, systems that fuse multiple data streams may require an understanding of the raw data that is hidden from the user. Without confidence about the source of data, users may dismiss the results as mysterious; and without proper skepticism in the limits of the data, users may place unwarranted trust in the results. Understanding the non-trivial computations and algorithms a system uses is also critical to allow confident tool use. For example, a global positioning system (GPS) navigation system performs a "black box" optimization that is often opaque to users. Whether a user trusts the system's chosen route can depend on their own understanding of the geography, traffic patterns, time of day, and so on. Part of gaining expertise in a system is learning to understand how underlying algorithms work, and if the system often produces unanticipated output, a user might begin to lose trust in the system. Yet part of the need for such systems is because people are unable to perform the processing on their own. This conundrum is a prime motivation for the EUG, as the EUG can help identify the boundary conditions where the result of the tool should be given less credence than others. Finally, output and display can be both an inhibitor and facilitator of tool understanding, trust, and adoption. Visual displays often rely on the human visual system's ability to make sense out of patterns, which can be helpful, but can also be misused (as argued by Tufte, 2001).

3.1.2 What Does an EUG Consist Of? The goal of an EUG is to compress many learned lessons users face along the pathway to expertise. These include experiences that both show how the target system can be used successfully; and experiences that illustrate the boundary conditions of the system, expert workarounds, and typical errors those users might make. Descriptions of some of these components appear in Table 12.

Table 12: Types of Learning Modules that can Comprise an EUG

Name	Description	Rationale
A wizard or walkthrough	Gives generic advice on how to use the tool properly for a new problem	After watching step-by-step instruction, many users are lost when it comes to repeating steps on a new problem.
Forced-choice scenarios	Presents a scenario, and a comparison of two alternative solutions	Shows positive and negative examples quickly with low overhead
Troubleshoot model	Presents a scenario with a specific problem, error, or misinterpretation, and allows learner to discover the problem.	Errors highlight boundary conditions and problem areas; this provides experiential support of that situation and its resolution.
Make error and show learner how to detect and avoid it.	A scenario is designed that compels learner to experience common error representing a boundary condition; resolution provides understanding of where the boundary is.	Novices are often not aware of their errors immediately; this helps them experience the errors and identify ways to avoid them in the future.
Give assignment; see expert solution.	A scenario is presented with adequate description, and the learner builds a model, then compares to expert solutions.	For modeling tasks, a tightly-scoped scenario can show range of valid solutions to problem. Multiple valid solutions can be valuable.
Give assignment, show steps expert took to solve problem	A scenario is presented with adequate description, and the learner builds a model, then watches the steps an expert uses to build model (with rationale)	In many cases, the individual steps taken are as important as the final result.
Work a problem with tool to answer a question about how the tool works	Non-transparent functions of the tool can be learned by posing a question about operation to the learner, and having him or her determine the answer via tool use.	Expertise develops with experiences where user discovers how a tool accomplishes some task or "thinks about" some problem.

3.2 Methods

As part of this effort, we examined two intelligent software tools in order to develop concepts and methods for creating an EUG. These include JCAT and NOEM (National Operational Environment Model). AFRL is developing both tools and their aim is to assist analysts in different types of analysis tasks. Our primary work centered on JCAT from which we developed a useable EUG concept. We intended our study of NOEM to take the lessons from JCAT, apply them to a new tool, and further hone our data elicitation skills for in-progress tools without any recognized experts.

3.2.1 Data Elicitation Methodologies for JCAT. We selected the JCAT tool based on consultation with our research sponsors, chose it from a set of possible tools for several reasons (see Figures 9 and 10). First, the tool computes likelihoods on causal networks, which had a strong relationship to the Causal Reasoning strand of the research effort. Second, AFRL developed the tool, and had some advocates and users within an intelligence group, to which we had access. Third, there was a perception within AFRL that users of the tool could benefit from the type of support that an EUG could provide. These advantages were balanced with several disadvantages of selecting the tool; the most critical was that JCAT had a small user base, with very few users who might qualify as having any real expertise in the tool. This limited our ability to capture incident-based use cases that we might have used to help identify learning objectives.

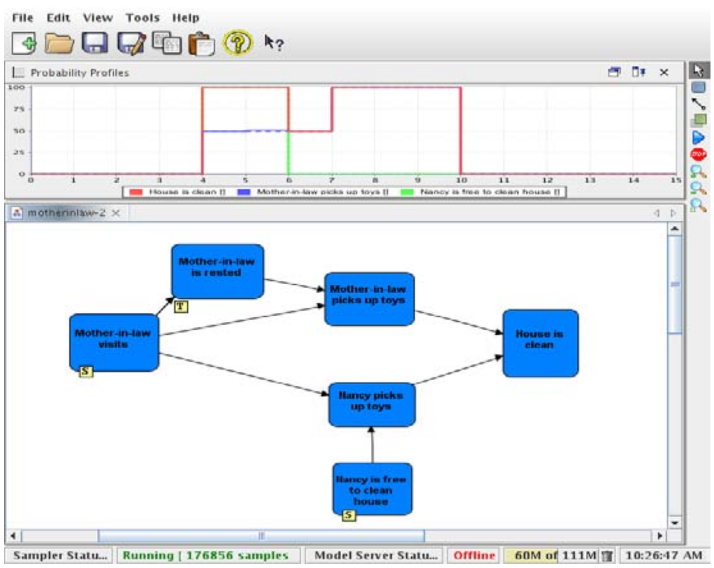


Figure 9: Screenshot of JCAT, Illustrating a Model (Network at Bottom) and Probability Profiles of Different Events (line graph at top)

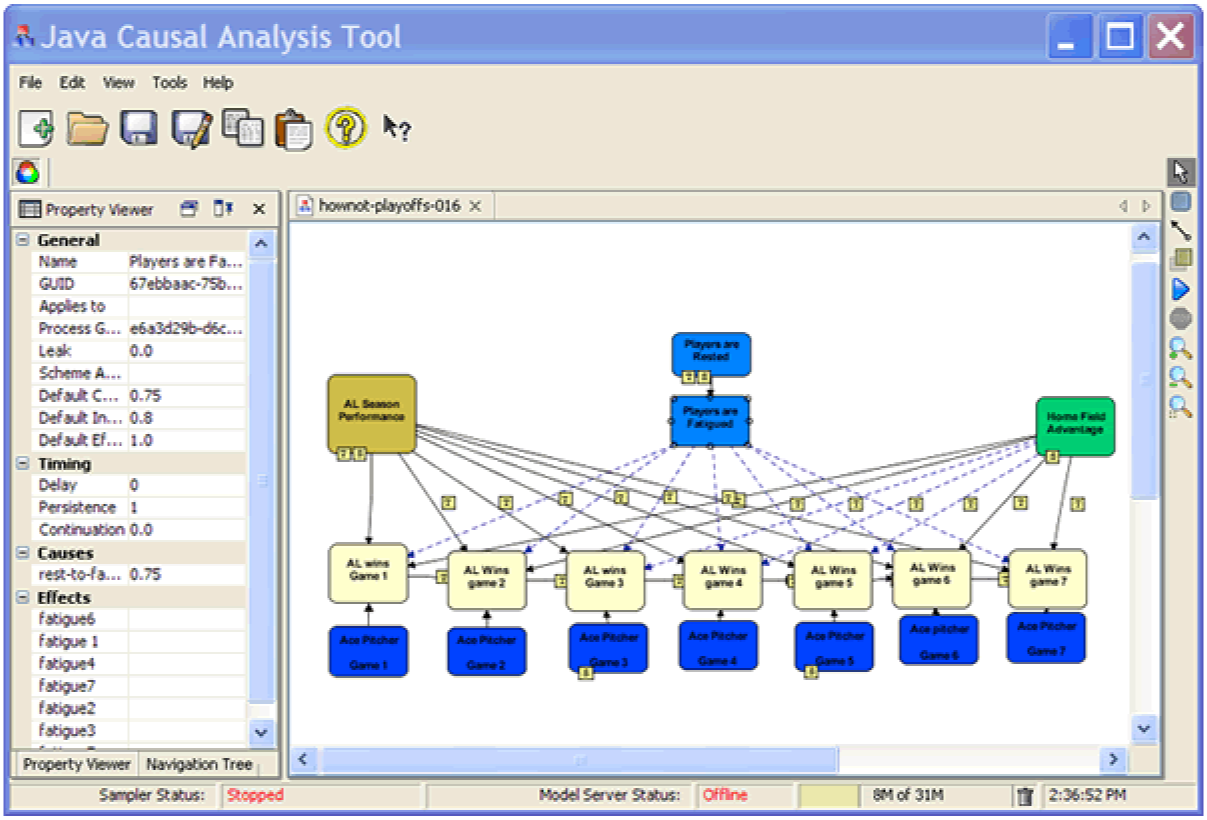


Figure 10: Screenshot from the Java Causal Analysis Toolkit

We used a number of sources to understand JCAT and identify learning objectives for an EUG. These represent both traditional and non-traditional data collection methods that can be useful in developing an EUG.

These included:

- *Examination of web-based forums.* The JCAT developers host a web-based user forum that includes questions and answers regarding the use of JCAT. In the case of JCAT, the forum was not extensive, but still offered a small set of questions and answers that represent areas users have trouble. Other similar sources that could be useful for developing an EUG for other tools might include user mailing lists, frequently asked question (FAQ) documents, and bug reporting databases.

- *Interviews with system's developers and trainers.* In this project, we had minimal access to actual JCAT users, and it is likely that even those who had used it occasionally have not developed well-practiced routines for solving typical tasks. This might be true for a large range of intelligent software tools, which an individual user might use only rarely. In lieu of a broad set of users, we conducted interviews with the JCAT development team, which included mathematicians, software developers, and individuals who had been responsible for training others in the tool. In the case of JCAT, the developers were also typically practiced users of the modeling tool, and so could provide us useful guidance about the tools intended use, and especially their mental model of how the tool operates.

We note that interviews with developers have to be used with some skepticism, in the context of other information. One thing we noticed is that because they understood the system from the underlying simulation engine, they found it difficult to understand some of the problems novices had trying to grasp the system from through the user interface alone. Furthermore, they had tendencies to think about real-world situations within the vocabulary of JCAT, and so found it difficult to identify situations the tool would not be appropriate for, even recognized how their reasoning had been captured by the tool. We would not be surprised if this type of phenomenon would occur for any advanced user of the tool. Nevertheless, it is remarkable how different a novice's experience can be from the experts: a novice may have a hard time thinking about a problem that a tool such as JCAT is useful for; while an experienced user has a hard time thinking about a problem the that the tool would not be useful for.

- *Examining models experts and novices developed.* We were able to obtain approximately 25 JCAT models that had been developed in a number of contexts, from by-products of novice training, to models developed as part of other training, to models pointed out by the developers as exemplary uses of JCAT. Examining these models showed many of the typical practices of experts, as well as errors of novices, and sometimes even errors of experts. Of course, not all tools have such useful by-products of the tools use, but when possible, such documents generated with the tool can prove valuable.

- *Interviews with experienced users.* Our original expectation in developing the EUG concept was that we would gather information on expert users via incident-based task analysis interviews. When we decided to pursue JCAT, we did so realizing our access to experts would be minimal. However, we were able to conduct one such interview with a moderately experienced user independent from the development team. For envisioned tools or beta-level software, these experts may not really exist, and we believe that using a number of other complementary data collection methods can still produce a useful information for an EUG.

- *Observations of 'first 20 minutes.'* Because the EUG is intended for relatively inexperienced users, it was important to observe several users exploring the system for the first time. We observed three such users in order to understand some of the obstacles someone experiences using the tool for the first time. This type of exercise can uncover many detailed problems with the user interface that users who make it passed the first session learn to work around. This can be useful for the developers, but much of that type of problem is not appropriate for an EUG: if there are problems with the labels on buttons that should be something that is fixed in the software rather than documented in a user guide. However, these observations can provide detailed information about the novice's mental model of the tool. For example, because the main process of developing a JCAT model is drawing box-and-arrow diagrams, one novice user began treating it like other similar tools he had used previously, essentially creating a concept map.

- *Diligent effort and detailed notes as the EUG Developer learns the tool.* One important source of information for the EUG came from one member of our research team learning JCAT, and taking detailed notes about issues, misconceptions, and problems as they arose. This approach may not be possible for all tools. For example, tools embedded within a physical system may offer only limited access to the EUG developer. In that case, one would probably need interviews from a wide range of users. Having direct access to the software tool also facilitates developing EUG lessons, at every level from creating screenshots to crafting problems that highlight the lessons of interest.

- *Developed face-to-face training and tried it on a small group of new users.* As we developed the EUG, we had the opportunity to test some of the user support concepts on a set of new users, in a face-to-face training. This was valuable as it allowed us to identify specific conceptual challenges faced by novices, and allowed us to test some of the EUG concepts and get direct immediate feedback. One of the lessons we learned in this exercise is that although showing examples of improper usage of the tool can be valuable to help identify boundary conditions, it must be balanced with examples of proper use, so that training does not become demotivating.

A comprehensive list of possible learning objectives for a JCAT EUG is described in Table 13.

Table 13: Basic Learning Goals of a JCAT EUG and Sources Used to Identify the Goals

Basic Objectives	Descriptions	Sources
	Representation and Modeling	
Minimal requirements for valid model	Nodes, mechanisms, signals, and triggers are necessary	Novice models
Causation does not imply timing delay	JCAT allows causation within time steps, without expected delays	Analysis of tool components
Role of conditional probability	Confusion in novices of whether it is for outgoing links or incoming links	Novices users
How to model event drivers	Leaks vs. external event versus scheduled event	Novice errors
Leaks are ‘unmodelled externalities’	Leaks should be reserved for ‘other causes’	Discussions with developers
Trigger Nodes	Model should be ‘driven’ be trigger nodes; external factors in the model	Developers/instructors
Network causal metaphor	Mechanism vs. signal: a signal is like an electric current in a wire	Interview with developer/trainer
Timing	Basic survey of the timing capabilities of a single node	Analysis of tool components
Basic Network Concepts	Nodes are EVENTS; links are causes	Interview with developer/trainer
Causal Loops	Tricks are needed to make a proper causal loop	Analysis of tool components
Proper definition of an event	A node can be mis-defined to represent event you are not interested in	Developers and model building
Causal versus Inhibitory impacts	Related but distinct from the inverted signals (above)	Analysis of tool components
Creating growth	At least two methods: invert a decay, or add persistence to a node/leak	Novice users
Level of Abstraction	How to choose the ‘right’ level of abstraction	Developers and novice errors
Does model match user’s concept?	Does representation in tool match what the modeler believes it to be?	Discussion with developers/users
	Data Generation and Data Handling	
Probability Estimation Errors	Overestimation, hard/easy effect, use of heuristics & biases	Academic Literature
Testing whether errors matters	Tricks for testing sensitivity of outcome to variations in inputs	Developers/instructors
Need to understand modeling domain	Domain understanding avoids many mis-steps	Developers; novice observation.

Extent of match between model and real-world process	Models necessarily are not identical to the process they represent; a user should understand matches and mismatches	Discussion with modelers
Inability to estimate probability	Inability to estimate probability is a hint that the model is framed poorly	Developers and trainers
Understanding Computations and Algorithms		
Timing overrides everything	Putting a scheduled timing event overrides external influences	Developers; model-building
Nodes really are events	It is an error to interpret a node as something other than an event	Developers and users
Cautions for multiple paths	Care must be taken to produce the correct intended effect	Discovered while using tool
Role/Use of sampler	Basic description of how it works	Observation of novices
Probabilistic events vs. causes	Identical outcomes, but making wrong choice can create problems	Developers; model-building
Meaning of OR	Basic notion of a 'weakest link' model	Analysis of tool components
creating an AND gate	Combining multiple signals to form an "AND" rule	Analysis of tool components
And & Or together	Non-intuitive effects when and/or concepts are combined	Discovered while using tool
Common causes of events	Related to problem/deter events/causes	Developers and model building
Use of Evidence	Important feature of tool that creates Bayesian logic	Analysis of tool components
Gates with On/Off logic	Use of inverted signals, and how they can create exclusive logic	Analysis of tool components
Conditional Probability Logic	How events are influenced causally by other events.	Observation of novice user
What-if games	Ways to 'pulse' a model to discover ways to modify it.	Interviews with developers
"Sopranos model" misconception	Misinterpretation of probability profile to be probability distribution	Developers and users
Output, Display and Visualization		
Basic Network Concepts	Nodes are EVENTS; links are causes	Developers and trainers
Networks JCAT is not good for	PERT, playoff charts, social networks, org charts, etc.	Developers and users
Probability Profile	Meaning of probability profile, distinction between "when something happens" and "how likely is something to happen"	Interviews with user; discovery via model building

3.2.2 Data Elicitation Methodology for NOEM. Following the development of the JCAT EUG, we turned to another tool developed by AFRL called NOEM (Figure 11). This tool was designed as a detailed model of the infrastructure (transportation, energy, etc.), economy, population, and government of large-scale (nation-level) entities. NOEM was somewhat less mature than JCAT, in that the tool was still in development and so had few real users outside the development team.

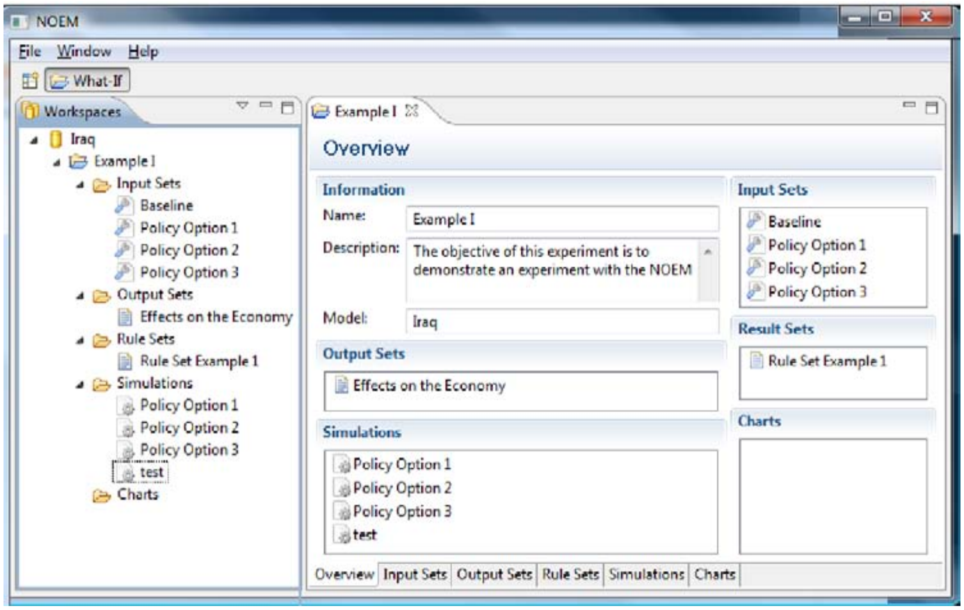


**Figure 11: Screenshot of the NOEM Tool,
Showing Basic Options for Creating Nation-Based Simulations**

Our primary elicitation methodology was a set of ten user observations and interviews that occurred jointly with a set of usability studies being performed by another research group. These studies involved both complete novices who had never seen the system before, and system developers, who had a range of experience with the tool. Some of the developers were responsible primarily for developing low-level models of different infrastructure components, and had in fact never before seen the NOEM interface. Other developers were responsible for the large-scale tool, and were perhaps the closest to experts in the tool that existed. We also obtained a working version of the NOEM tool, which two members of our research team explored enough to gain some reasonable competency, taking notes about misconceptions and problems along the way. Finally, we also had access a substantial number of NOEM user guides and documentation, which provided us some additional insight into the tools intent and purpose.

The most interesting insights from this came from contrasting the type of tool NOEM is with the type of tool JCAT is. Although both tools are in some sense modeling tools, the end users of the tools play very different roles. For JCAT, the user actually constructs the model, and the model serves as a device to enable structured probabilistic reasoning. For NOEM, the anticipated end user would have fairly little control over the structure of the model, and their main role is essentially to explore the parameter space of the model to understand consequences of different policies. The actual modeler in NOEM is someone who hopefully has expertise in the domain and region being modeled. The end user‘s job is essentially to craft a policy (perhaps increase policing or decreased wages), and look at the outcome over several years of that policy change, in comparison to leaving it the same.

Some of the basic lessons learned from these observations were:

- *Timescale.* Novices had little idea about the appropriate timescale for drawing conclusions about a NOEM model. This time-scale is somewhat tied to the types of inputs available, which are intended to represent policy changes. We inferred from interviews that the proper time scale for a nation-level model is at least several months, but probably two to five years. Although the simulation can make predictions about immediate consequences, and can make long-term predictions, both of these extremes are limited.

- *Underlying Dynamics.* Users with greater understanding of the underlying system dynamics model had intuitions about how processes rise and fall to equilibria. NOEM is primarily a system dynamics model, which involves simulating stocks-and-flows via numerical approximations to differential equations. Thus, it is at its core a hydraulic model, with some stochastic event simulation when appropriate. A relative expert might look at a rise to asymptote in some quantity as a signal that the system's initial conditions were wrong and the system was reaching an equilibrium; a novice might see the rise and try to attribute it to the impact of another variable.

- *Variability/Randomness.* Along with the smooth dynamics that are typical for NOEM, there are a number of time-based stochastic events. For example, power stations could be either on or off, and the system would frequently model power outages stemming from disrepair or sabotage. When a power station went down, the local economic output in its region would also diminish for several days or weeks. This revealed a counter-intuitive finding: when novices saw these dips in economic output, they inferred that something was broken, because they did not know the reason for the dips. Experts understood both the source of the dips (power outages) and the fact that power outages were discrete random events, and so the dips became a signal that the model was working correctly. Furthermore, understanding the underlying stochastic model led experts to infer that the outages were representative of the types that could be expected, and not specific predictions about an outage on a certain date.

- *Geographical Scope.* The most well developed NOEM focused on Iraq, with most of the model focusing on Baghdad. We anticipate that some users might have different expectations about the actual or proper scope of such a model. One might find it difficult to justify a model of Iraq that does not incorporate Iran and Turkey, or that does not provide detailed geographical models of the outlying regions.

- *Modularity.* Although in a real society, areas such as health, economy, government, security, power, employment, etc. are tightly coupled, NOEM models these as distinct modules with tightly coupled variables within a module, but often much less interactivity between modules. Thus, if one is trying to find the source of some observed effect, the user cannot rely on intuition or understanding of the nation of interest, but must also think about the structure of the model. Parameters within a particular module have a much higher chance of affecting one another, and can be remarkably unaffected by variables in other modules, that one might assume are related. This may partly represent an incomplete model, but the notion must be understood because any model of a real-world process is necessarily incomplete.

- *Importance of Different Variables.* NOEM offers hundreds of variables that can be adjusted or monitored during a simulation. However, in example problems, users kept returning to a few core variables to change or monitor. The importance of these variables is not readily revealed in the interface, and the most central ones appeared to be both variables that were probably important for policy-makers, but also variables that were known to have impacts on one another.

- *Parameter Search Strategies and Workflow.* Our interviews revealed many issues with the NOEM interface that caused problems for users. Some were simple things like button labels, and others related more to workflow. We believe that EUG is not appropriate for teaching about little interface issues; on the other hand, issues related to workflow might be, and so we discuss them here. For example, one might want to increase or reduce a particular variable of interest, and want to identify other variables that could affect it. This might require a search through parameters, and NOEM does not currently support such a search well. Another example might be managing a 'campaign' of models that systematically change variables to show the combined effects of some set of changes. This type of task must be managed by hand, and experts may develop practices to streamline this type of process. Such workflow lessons could form part of an EUG on NOEM.

A detailed NOEM EUG was not developed for this contract. However, we believe such a tool could be a useful augmentation to the current user guides, to help provide novices a better understanding of how to use the tool, rather than just what the tool does.

3.3 Results: Types of EUG Lessons

During the research effort, we identified a number of distinct lesson types that can form an EUG. Brief descriptions of these follow, along with examples from the JCAT EUG.

3.3.1 EUG Lesson Type: Walkthrough or Wizard. Most software documentation will provide a detailed description of how to use features in isolation, but it is rarer for software user guides to provide instructions for how those functions should be used together to accomplish a goal. Sometimes, software will embed a set of operations in a wizard; usually a sequence of dialogs that walks the user through the set of operations in a task. These are especially useful for configuring the system; tasks needing be done correctly for the system to operate, and require some user decisions to be configured properly. This is also similar to the "Walkthrough" concept created to assist in many computer games: a guide that provides a sequence of operations that allow the task to be accomplished.

We envisioned an EUG walkthrough having the following properties: it should be embedded within a particular problem which should illustrate a sequence of abstracted steps that could be applied to a new problem. At each step, rationale for why that step is performed should be provided (or reasons why this step might be omitted). At the end of the walkthrough, the

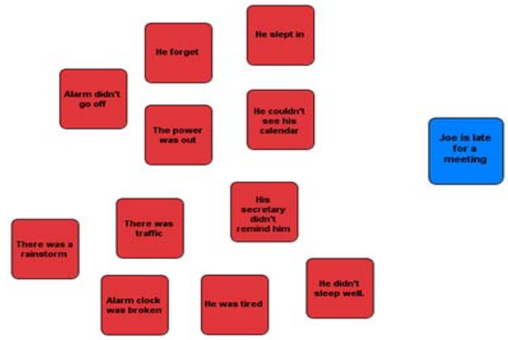
abstracted sequences of steps should be provided, so that the same steps could be applied more easily to a new problem. Finally, the developer should identify the boundary conditions on use of that solution: is it the recommended expert solution; or a simple novice solution.
We developed a wizard lesson for JCAT. For the walkthrough, we identified eight basic steps that we had learned from expert users were a good way to go about building a model. This method actually represented one of the two main methods experts described as possible methods, but was also modified to incorporate necessary steps that model builders did not describe but were necessary.

This lesson had eight steps (Figure 12), which were covered in eight screens of the EUG. The steps are:

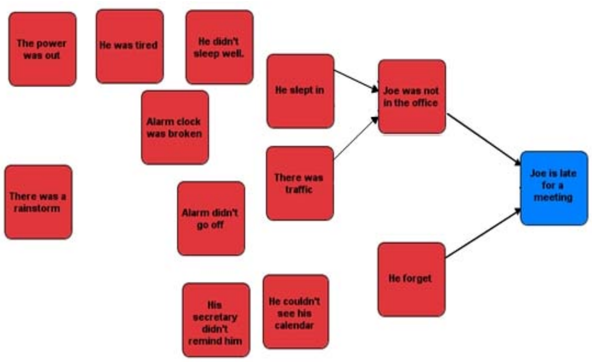
1. Identify final outcome node.
2. Create events that might influence this node.
3. Work backwards, adding nodes to the network causal structure, but do not set probabilities.
4. Identify trigger nodes: events at the edge of the network that drive the network.
5. Schedule the trigger node events with appropriate probabilities.
6. Move forward in the network, setting conditional probabilities and testing whether a node has any probability of occurring.
7. When assigning conditional probabilities, notice when multiple causes have a combined effect that differs from their individual effects.
8. When complete, the outcome event should give an answer to the probability you wondered about at the beginning.

Step 1: Identify Outcome Node

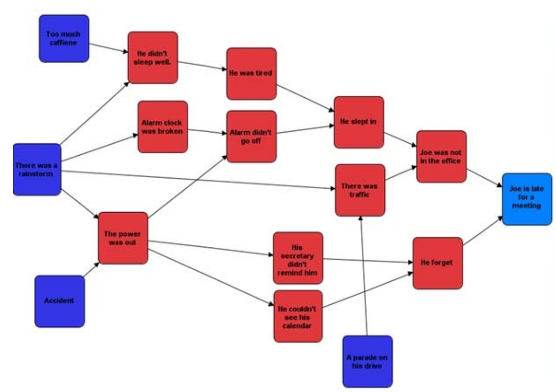
Step 2: Create events that might influence this node



Step 3: Work backwards connecting events in causal structure



Step 4-7: Identify trigger nodes; schedule events and set conditional probabilities and combination rules by moving forward in the network.



Step 8: Final probability gives answer to question.

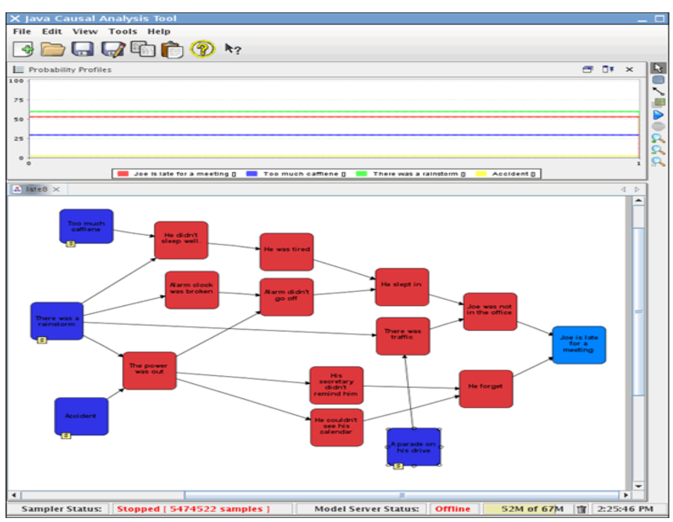


Figure 12: Step-by-step Walkthrough EUG Lesson

It is important to recognize that not all models need to be built in this way, but this is one structured method that can help novices build valid models.

3.3.2 EUG Lesson Type: Forced-Choice Comparison Sets. An important lesson that users of a new system must learn (and often learn continuously) is the tasks, situations, data sets, and contexts in which the tool is well suited for, in comparison to those it is ill-suited for. An expert will not only know situations in which it is inappropriate, but also have a good idea of where the boundary condition between appropriate use and inappropriate use is. Consider an office suite user who wants to create a table of information for a word processing document. In some situations, using the built-in table functions within the word processor would best serve the user, such as in Table 14:

Table 14: Built-in Table Function within the Word Processor

Brown Things	Red Things	Green Things
chocolate	strawberries	grass
cardboard	stop signs	kiwi

In other situations, it would be more efficient to create a spreadsheet containing the information, and import this table, as in Table 15:

Table 15: Spreadsheet Created for Import

	Type 1	Type 2	Type 3	**Total**
Group A	2	23	55	**80**
Group B	2	33	42	**77**
Group C	1	31	55	**87**
Total	**5**	**87**	**152**	***244***

There may be several boundary conditions that might help the user choose which one to use; one involves whether the tabular information is numerical and requires some computation, as in the second table. Most users of office software have learned boundaries such as these through experience. However, software developers and vendors are reluctant to highlight situations where their tool is not appropriate. For example, Microsoft provides a 30-page interactive tutorial on table design[1] including examples where numerical data are formatted, but never once mentions that another method for creating a table is to import an Excel selection. This just exemplifies the lack of explicit discussion of boundary conditions in user guides and documentation.

[1]http://office.microsoft.com/en-us/word-help/tables-i-create-and-format-basic-tables-RZ001200716.aspx

Our concept for helping to identify boundary conditions within an EUG involves what we call a "contrast set" or a "forced choice scenario." These are intended to be embedded in example problem, and to show two cases of possible interpretations, outcomes, input conditions, and so on. After describing the scenario, the user is given a choice between two cases, and they must chose which is correct. The choices are designed to straddle a critical point on a boundary condition.

As we developed the EUG for JCAT, we identified a number of different variations on this forced-choice scenario. One thing we learned is that they can be fairly brief, and it may be more efficient than developing complex use scenarios that the user then explores with the tool. We also identified a number of variations of the method which help support different types of goals. These variations are shown in Figure 13. Each variation involves describing a single stage in the modeling process, and having the user choose between two alternatives at another stage. All of these options include a description of a model, either as a problem description or as a pair of options.

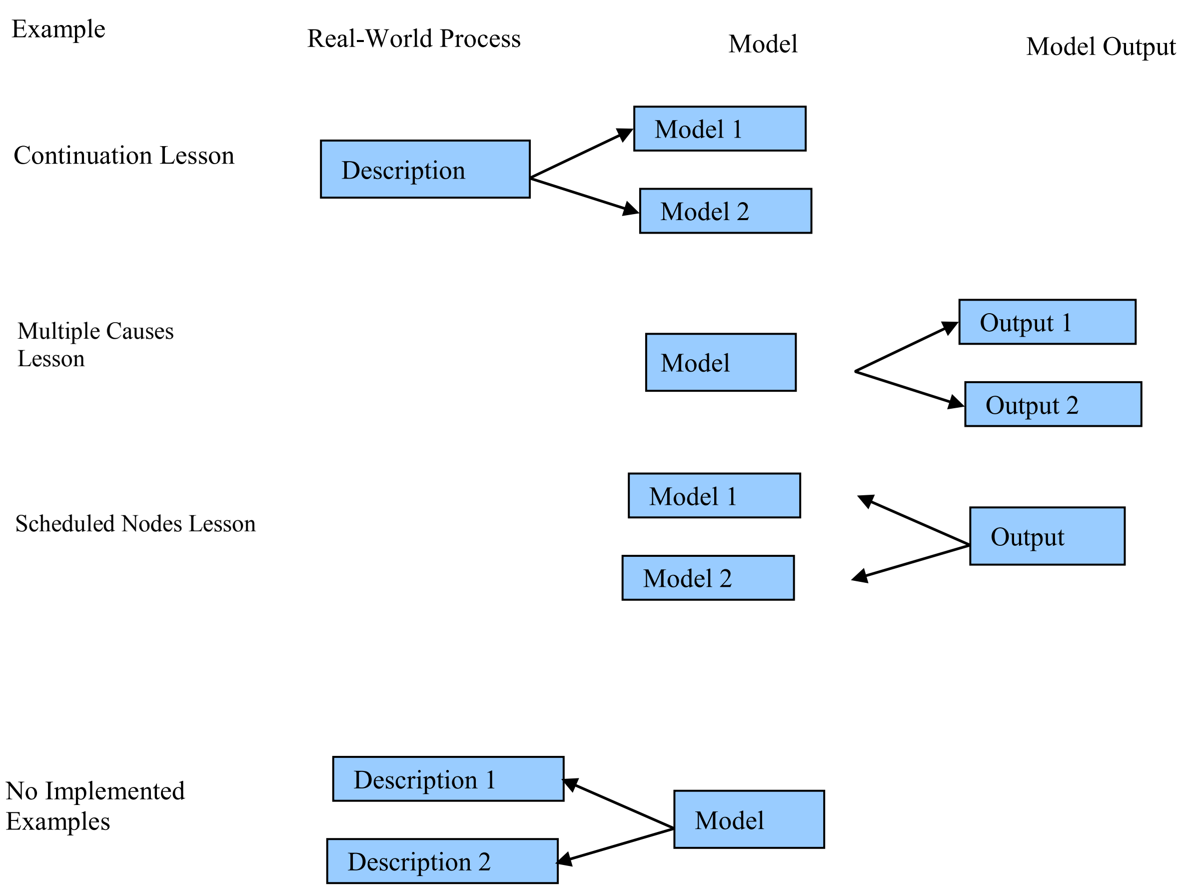


Figure 13: Variations of the Forced-Choice Scenario. Illustration of Four Different Forced-Choice EUG Module Types

Example 1: The Role of Scheduled Nodes

A common novice mistake we saw when developing the JCAT EUG was a misunderstanding that a model could only produce results if it had some intrinsic input nodes. These are sometimes referred to as trigger nodes, and they are typically created by scheduling them to occur using the JCAT Timing dialog. Although there are a few other (somewhat indirect) ways to generate input into a model, one can usually assume that if a model has no nodes that have been 'scheduled' to fire at a specific time, the model won't work. In fact, if a root node on a chain of events has no inputs and it has not been scheduled, the entire branch is essentially a dead, and will have no impact on the model as a whole.

Whether or not a node has been scheduled can be seen within the display interface: it shows up as a yellow 'S' note on the lower left of a node. We observed more experienced users visually examining the network to identify the "S" nodes, which helped them diagnose whether a network or branch of a network was being used. This lesson thus helps support two learning objectives regarding proper use of the tool: 1) understanding the difference between a trigger node and a non-trigger node; and 2) learning to identify these by visually inspecting a network, rather than digging down into dialogs.

The setup for this problem is deceptively simple. We show a screenshot of a network that has two nodes, and produces a single output (Figure 14). The user is asked to identify which input model produced the output, even though they do not show the underlying timing dialog that could be used to determine this. The only difference between the two nodes is the letter on the yellow annotation, with one indicating a scheduled trigger node (S), and the other a timing property (T). The goal is to help the user understand that the "T" annotation can be a critical display property to help predict what a network will do, and also to begin learning that an event node without a "T" will not occur on its own.

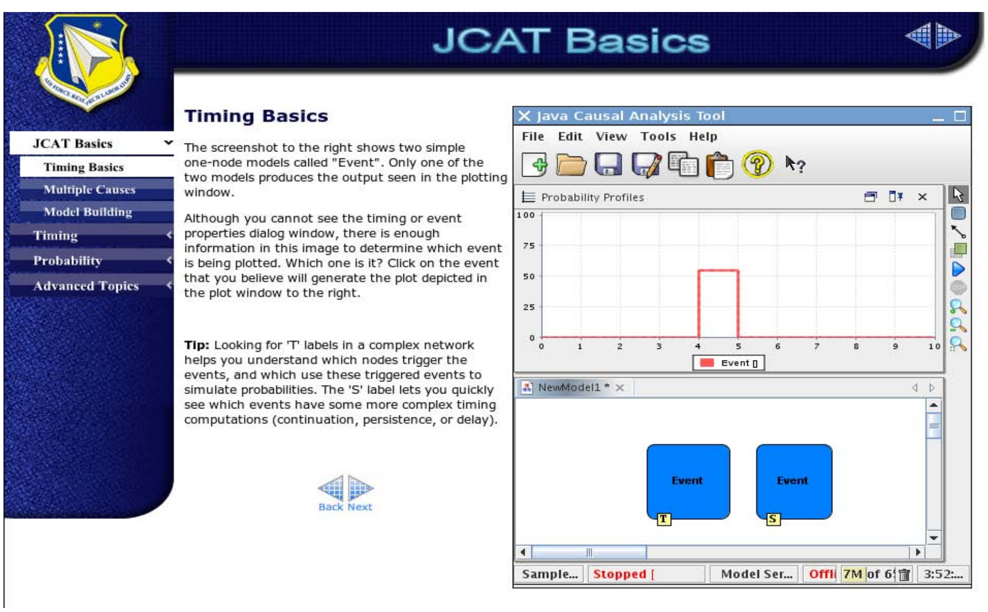


Figure 14: EUG Setup for the Role of a Scheduled Nodes Problem

In the JCAT EUG, choosing the incorrect option provides a short lesson on the meaning of "T" and "S" (Figure 15). Choosing the incorrect option provides explanation for why it is wrong, and directs the user to the correct option. Choosing the correct option provides a little explanation for why the option was correct (Figure 16)

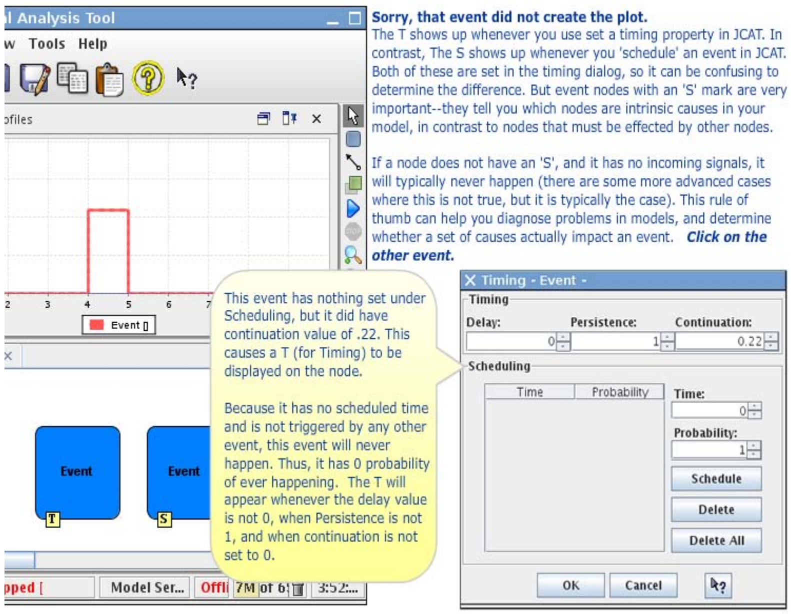


Figure 15: EUG User Feedback or an Incorrect Choice

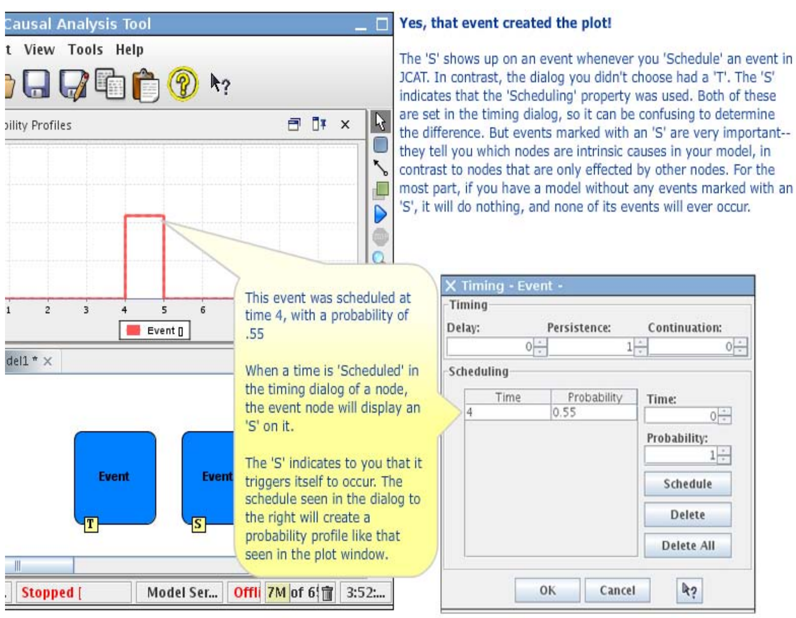


Figure 16: EUG User Feedback for a Correct Choice

This version of the forced-choice scenario focuses on showing a single output, and requiring the user to choose between two possible inputs that produced that output. Such a scenario could be useful in other contexts, especially in developing an intuition for how some complex software system should behave.

Example 2: Multiple Causes

Another use for the forced-choice scenario is to show a single input with multiple outputs, and have the user identify which output was produced by the input. This purpose is slightly different from the earlier one, because it could be especially useful for making predictions. That is, by showing a single input configuration and forcing the user to choose between two outcomes, it enables them to think about the possible ways a system could transform data.

Initial scenario description is shown below. In this scenario, the primary objective is to help the user understand the difference between two ways of combining multiple inputs. By default, the incoming links operate according to an “or” rule: a single incoming causal event can trigger the effect; and multiple incoming events increase the probability of a later event happening. Thus, it works analogously to a chain, which will break when any of its individual links break. But some events only happen when multiple other causal events occur, such as an airplane failure because of engine failure. In the case of a two-engine airplane, both engines must fail before the plane loses power.

Figure 17 shows the forced-choice scenario, which shows a single network model with a basic description of its organization. When selected with the mouse, each node displays the relevant timing or properties dialog, which would be accessible in JCAT via a context menu. Two possible outcomes are shown, each one mapping on to one of the two rules for combining probabilities. The user must select the correct outcome; if he or she chooses incorrectly, appropriate feedback is provided; otherwise an explanation for why the choice was correct is given.

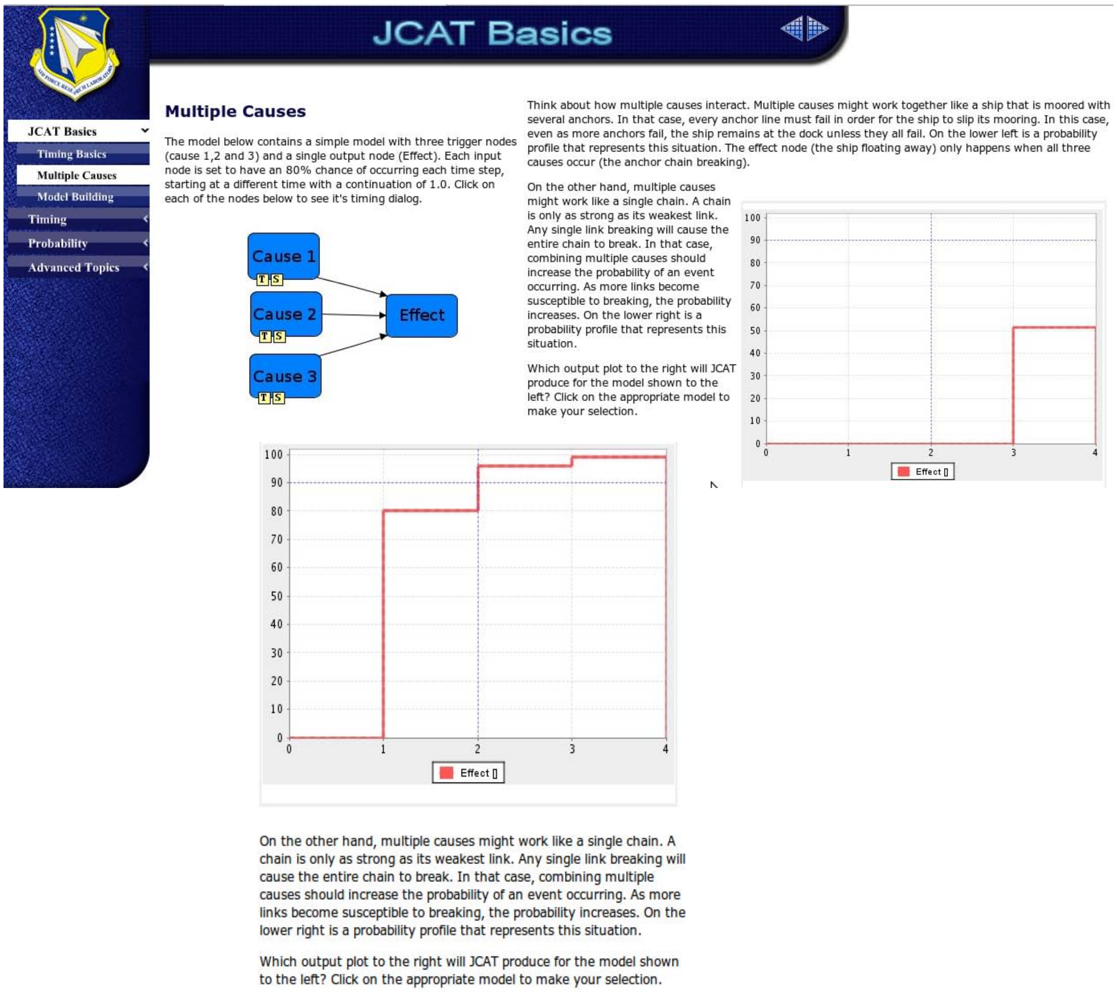


Figure 17: EUG Forced Choice Setup for the Multiple Cause Problem

3.3.3 EUG Lesson Type: Troubleshooting/Induce ERROR. Another type of lesson has the following rationale: When a novice is learning to use a complex tool, some of the most critical learning events come when an error is made and the novice must reconceptualize his or her mental model of the tool in order to overcome that misconception. Without being encouraged to do so, a user will stumble upon such problems only occasionally. Therefore, we conceived of several versions of 'troubleshooting' scenarios, in which scenario is presented in which some error or misconception has been made, and the user must explore the tool in order to identify what the incorrect assumption was that led to the problem (Figure 18).

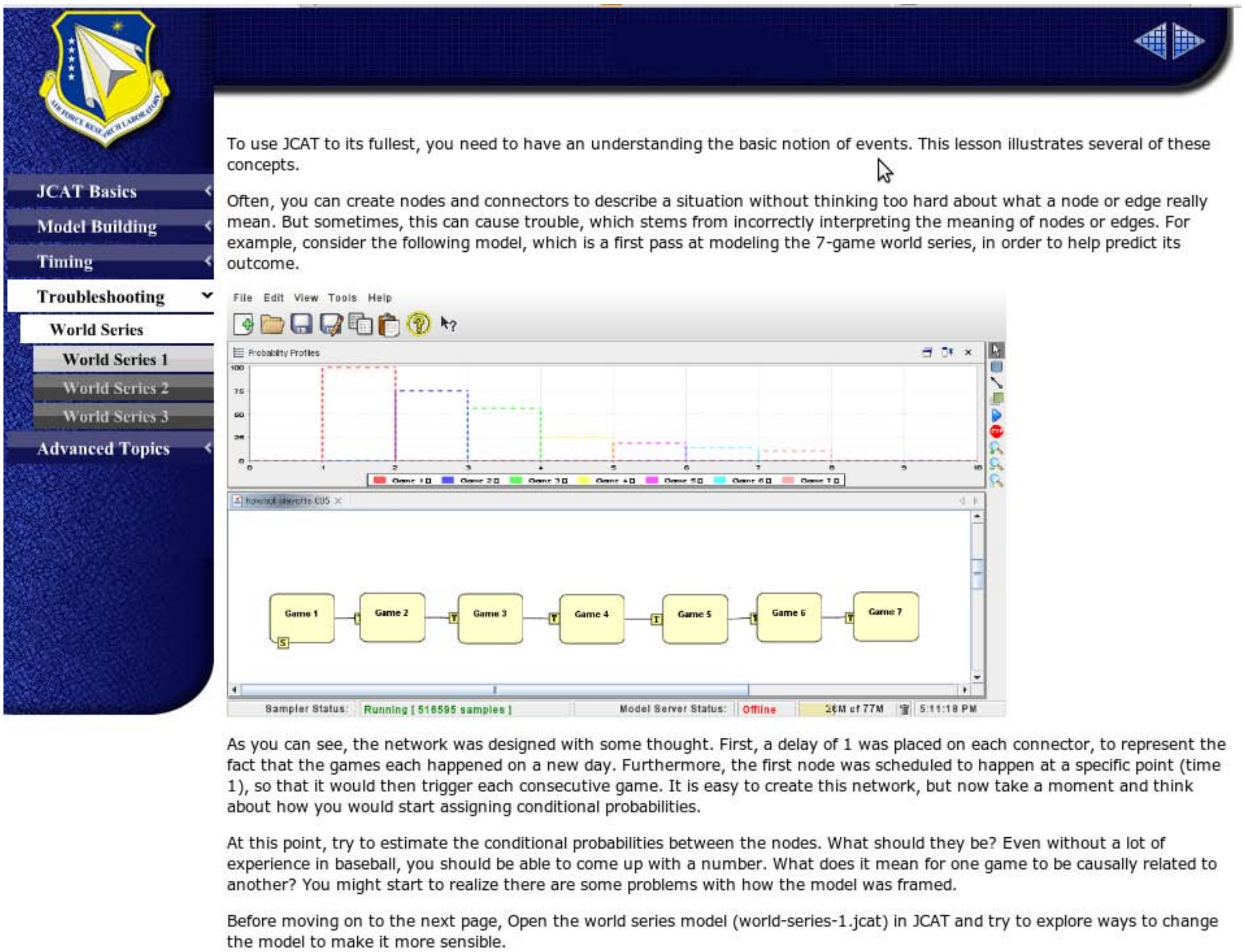


Figure 18: EUG Setup for the Troubleshooting/Induce Error Lesson Type

A variation on the troubleshooting scenario does not simply present a scenario in which an error has already been made, but rather tries to induce the learner to make the error himself or herself. This type of lesson is more difficult to produce but possibly has a better pay-off, because it better replicates the experience of making the error. However, this type of model may not work because it violates the assumption of the user that the user guide is not attempting to trick them. For example, in order to induce the user to make an error, the user guide will need to mislead the user into conceiving of a problem in a particular way, so that they make a miss-step. The resolution might be to reframe the given problem in a different way, such that it challenges the assumptions of the problem, and this type of scenario may be seen as misleading, even if it teaches a valuable lesson.

We tried these types of lessons with novice JCAT users and learned that these types of lessons should be used with caution. It is important to see the errors, but if one sees too many errors, it gives the (possibly mistaken) impression that the tool is very easy to use incorrectly even when the scenarios need to be contrived carefully to illustrate the mistake.

3.3.4 EUG Lesson Type: Work Problem and See Solution. A final lesson type we identified provides the user with a basic description of a problem which they must create a model for, and then shows them one or more models developed by others. These types of lessons have less of a chance of addressing a particular learning objective, but can be valuable in giving the user experience creating models and looking at alternative approaches. Although just the act of comparing one's own model to the model of others can be useful, it is also useful (but more effort) to provide more detailed explanation of why certain decisions were made, so that the user can follow this process in the future. Showing multiple solutions can give a feeling for the range of approaches available within the tool.

Our prototype EUG contains one lesson specifically implementing this idea. In this lesson, the user is asked to create a model of the risks that might predict high blood pressure. The solution is shown in Figure 19.

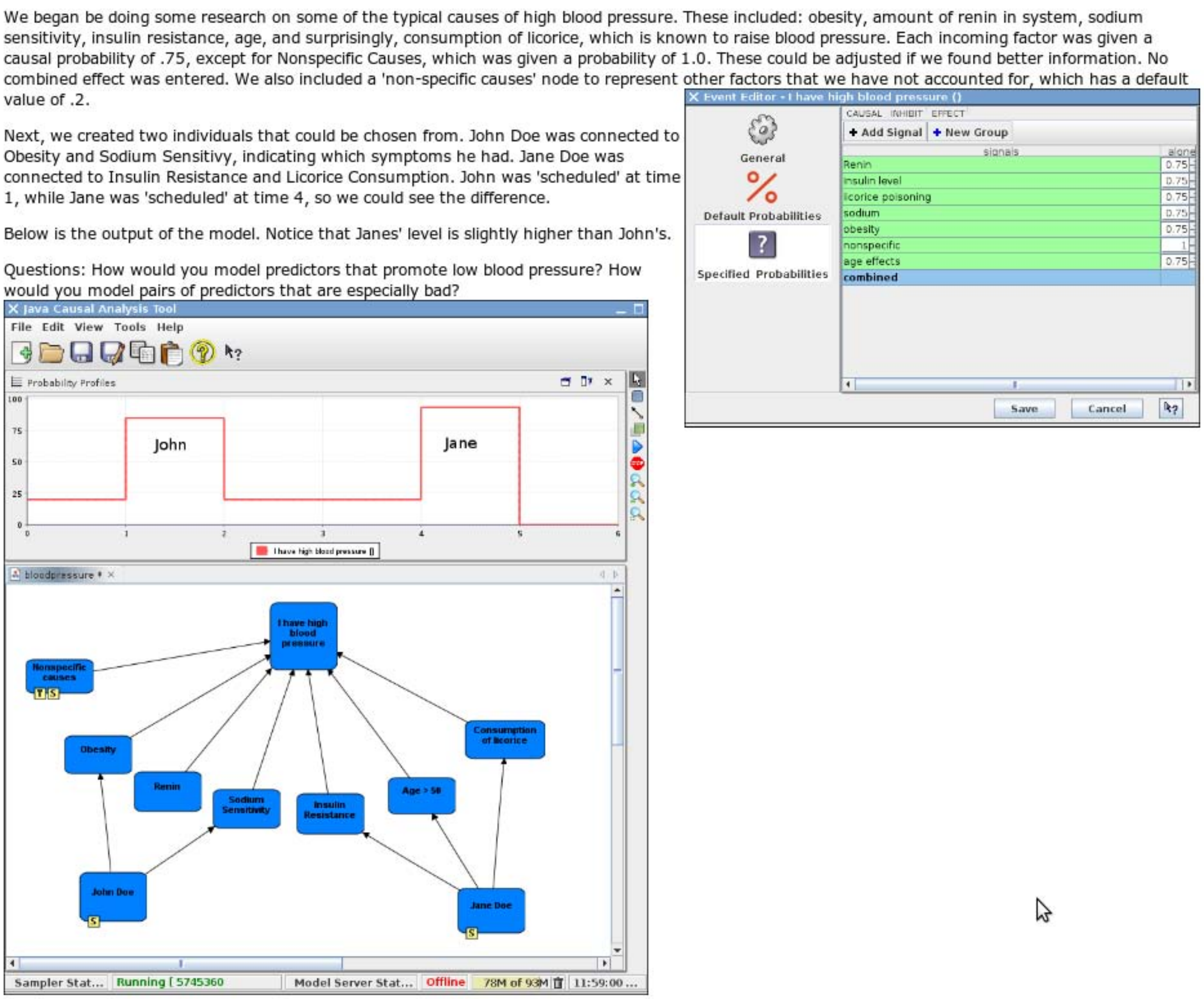


Figure 19: EUG Setup for the Work Problem and See Solution Lesson Type

3.4 Discussion: Boundary Conditions of an EUG

Our approach in this project has been to identify the methods for creating an EUG. Along the way, we developed useable training for one system (JCAT) and made important first steps toward developing training for a second system (NOEM). Although our concept of the EUG has changed because of this research, it is important to understand what the bounds of an experiential user guide are. Does any interactive training qualify? Therefore, just as the EUG attempts to use different modules to illustrate the boundary conditions of a tool, we will try to identify some of the boundary conditions of an EUG.

3.4.1 An EUG is for Non-Trivial Intelligent Software Functions. Experiential training has proven effective in many contexts. One of the recurring discussions we had during this effort is in determining whether some type of lesson is applicable for an EUG. Our intent is to focus on tools and functions of tools that perform non-trivial reasoning. So although one could use the lessons such as we developed to help a new user understand the location, meaning, and purpose of different buttons or menu, it is unclear whether the investment required for an EUG would be worthwhile in those contexts. For straight-forward software functions, the scenario-based learning may also be inefficient. A user may simply want to look up the process in a detailed traditional help system, and not have to wade through scenario-based instruction to discover the answer. However, for more complex functions, which require the user to have an accurate mental model of some internal process, the EUG is more appropriate, because it can help embed the training in examples that are relevant and useful to the learner.

3.4.2 An EUG is not an Introductory Course Book. One answer that must be addressed when conceptualizing an EUG is what type of experience a user will have prior to picking up the EUG. For both of the systems we examined, there was little formalized introductory-level training material, which posed a dilemma. Ideally, users might receive classroom introductory training which would cover the basic function of the tool, and later use the EUG on their own to develop a better understanding of some of the more challenging concepts. However, realistically we could not be certain about the capabilities of EUG users, and so some of our lessons covered concepts that would have better appeared in those introductory materials. We suspect that this will be a issue that needs to be balanced any time an EUG is created.

3.4.3 An EUG need not be Web-based Training. Our JCAT EUG was implemented in the form of an interactive html and flash-based training platform. This has certain advantages, in that it can be deployed remotely, and offers freedom to use multi-media training in various formats. We believe, however, an EUG is primarily about the cognitive challenges that users face in complex software tools, and not necessarily about the mechanisms by which training is delivered.

3.4.4 An EUG should not Replace Usability Testing. Usability concerns are almost always a huge challenge for special-purpose software with small user bases that we anticipate the EUG being most valuable for. The process of developing an EUG can unearth many of these issues, but we recognize there is a distinction between the types of usability issues revealed by standard human factors user testing, and the types of problems the EUG should focus on. Usability testing is often predicated on the notion that the interface can be made easier to use. The EUG should support the types of functions that cannot be made easier, and will necessarily require experience and training to get right. By analogy, user testing might help one design how to design the ergonomics of new musical instrument, but it will never make the instrument so easy to play that even a novice can perform masterfully.

This EUG attempts to expose a user to many of the experiences that an expert will have over a period of weeks or months of use that allows them to understand the strengths and weaknesses of the tool. It is intended for software tools that perform non-trivial "intelligent" functions, and can ultimately help both novice and experienced users gain new and better understanding of their tools.

4.0 CONCLUSIONS

Our research supported our speculation that the standard views of causal reasoning are limited in their applicability to natural settings. The standard view is that people identify an effect they want to explain, then they start by nominating possible causes, they evaluate each of these causes, and they select the best one. This view is systematic, and it works in determinate situations where effects hold still, where causes can be specified and evaluated.

The standard view of causal reasoning contains a number of myths that result in confusion rather than clarity for natural settings:

1) There is a single cause and we can find out what it is. Thus, the detective work to diagnose what causes AIDS, what causes yellow fever, how birds migrate successfully, all result in clear and satisfying answers. However, in indeterminate settings, we will never know why Hillary Clinton lost the Democratic nomination in 2008 or why the South lost the Civil War, or why a specific project is going well or poorly. The most difficult, and usually most important, causal reasoning is done in indeterminate situations where there are not systematic ways to determine the "true" cause – because the notion of a "true" cause is misguided.

2) There is a single cause. This is a particularly pernicious myth because it conditions what counts as a satisfying explanation. In a determinate situation we can usually point to a single cause, but not always. For indeterminate situations, we usually encounter a causal field and our attempts to land on a single cause result in shallow explanations. Why did the U.S. military do better in Iraq starting in 2007? There was the surge, but also the over-reaching of Al-Qaeda in Iraq (AQI) when it began terrorizing its Sunni supporters, leading to the Sunni Awakening that turned the populace of Anbar Province against al Qaeda. Also, the military decision to reverse field and support the Sunni militias instead of trying to defeat them. Plus, the assassination of Abu Musab al-Zarqawi, leader of the AQI, which was part of a general view that the U.S. was now the "strongest tribe." No one of these factors turned the tide, and they are inter-related. Any attempt to single one of the factors out will be impoverished.

3) There is a clear effect to be explained. We found the "effect" sometimes morphed in indeterminate settings because the investigation into the cause resulted in clarification of the effect. Thus, we might want to understand why the U.S. set up a blockade in October 1962 after discovering the Union of Soviet Socialist Republics (U.S.S.R.) was placing missiles in Cuba. In trying to formulate an explanation we might realize that we are trying to explain how the U.S. chose to use a blockade, or why it was set at 500 nautical miles from Cuba, or what the policy was for letting some ships through but not others, or when the blockade started. Often, people may be explaining different aspects of an "effect" and thus talk passed each other without realizing it. We have developed an "effect ontology" to describe different aspects of effects that we seek to explain: factors in the mind (e.g., changes in a belief or an attitude) and factors in the world (changes in a world state, changes in a policy, changes in a course of action, the decisions/actions of another person, the absence of an event/effect).

4) A good explanation provides a single cause to account for an effect. Our research demonstrated a variety of causal formats, and we found that we could manipulate which type of format people prefer.

5) The goal of causal reasoning is to deepen our understanding. We found that there are two different goals for causal reasoning. The first is to gain a deeper understanding – but this is best achieved by adding complexity to the causal field rather than seeking convergence. The second goal is to support action, and here people do prefer simpler causal reasoning formats, such as chain reaction models or single cause explanations.

6) Once we know the cause of an effect we want to avoid we are on our way to preventing it. This myth applies in determinate, single-cause situations, such as medicine. Once we determined that mosquitoes transmitted yellow fever and malaria, we could start to implement prevention strategies. Once we determined that contaminated water led to cholera, we could impose sanitation measures to prevent further outbreaks. However, when dealing with indeterminate, multi-causal situations, the picture is not so easy. We may have a chain of events, or more likely, several chains, or even more likely, several chains that intersect. Interdicting any one of the chains might do the trick. Or not. We can speculate on why Hillary Clinton lost the Democratic nomination, but we will have a list of reasons, at best, and we will lack guidance on what she could have changed to alter the outcome. Or consider the friendly fire incident (Snook, 2000) in which the U.S. shot down its own helicopters. Snook provided a diagram of all the causes he identified (Figure 20). At first, this complex diagram appears daunting but actually, interrupting any of the chains might be sufficient – making sure that the helicopters were filing some sort of flight plan, or ensuring that the helicopters were squawking the right IFF codes once they crossed into Iraq, or alerting the Airborne Warning and Control Systems (AWACS) Weapons Directors to be on the lookout for U.S. helicopters that need to be reminded about the identification friend or foe (IFF) codes, etc. Any of these or a few other actions would do the trick. Thus, the friendly fire diagram is largely an AND diagram, so reversing one of the critical nodes will be sufficient. The Hillary Clinton failure is an OR diagram, and so reversing any of the individual nodes might not achieve much.

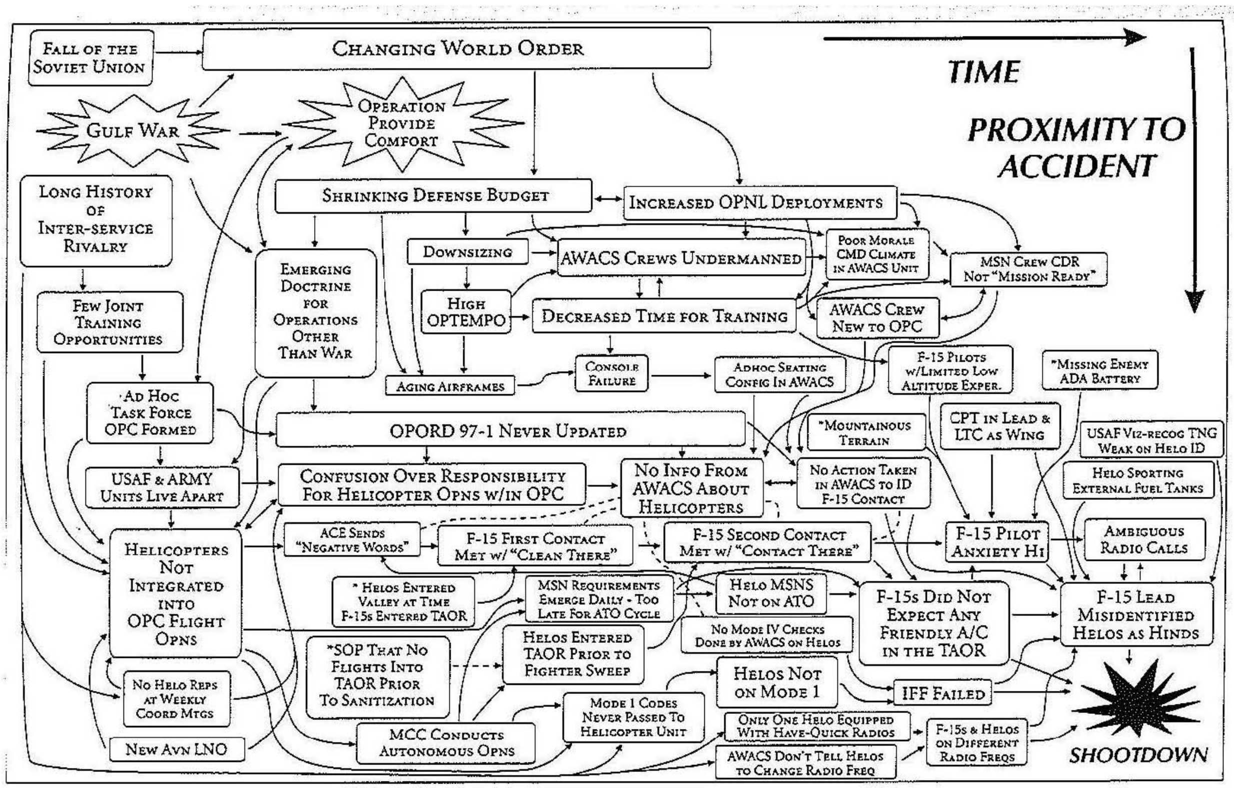


Figure 20: Causes Leading to Friendly Fire Incident in which the U.S. Shot Down One of Its Own Helicopters (Snook, 2000)

Taken together, the limitations of these myths explain why approaches such as Six Sigma and Kepner-Tregoe have such limited value. They assume deterministic settings in which careful diagnosis can pinpoint critical flaws. When these kinds of analytic methods are applied to complex and indeterminate situations, they are much less useful.

We now have a more nuanced model of causal reasoning in natural settings, as shown in Figure 21. Compared to the Butterfly model, we have decided that the single cause explanation of providing an abstraction is much more sophisticated than the other single cause formats (an event, a decision, a force). In addition, we find no mention in the literature of the abstraction as a causal reasoning format, and yet in our samples the abstraction was often generated.

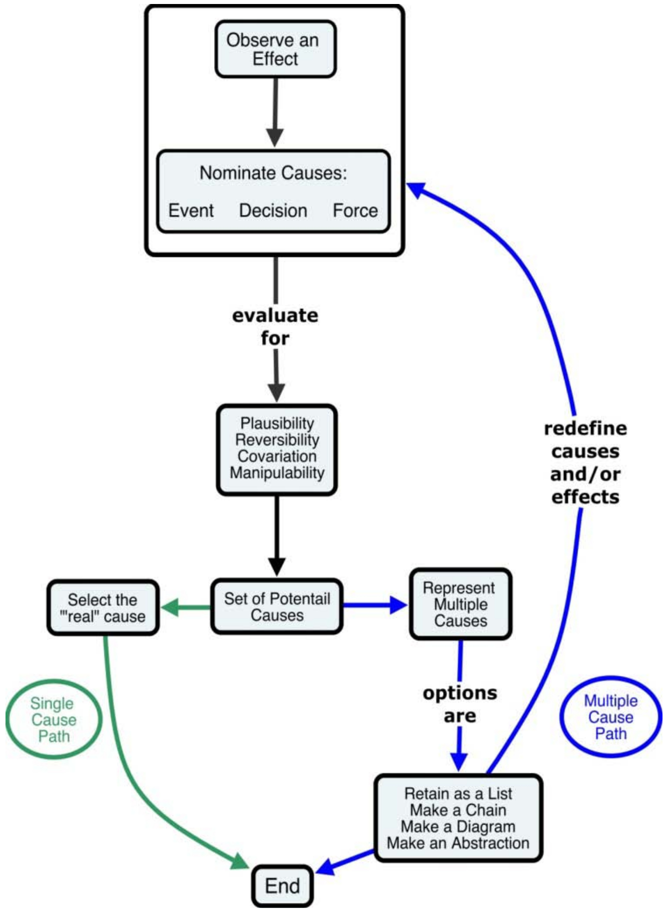


Figure 21: Model of Causal Reasoning in Natural Settings

We have also generated some approaches for applying the lessons we have learned about naturalistic causal reasoning.

One class of applications is about guidance to help people do a better job of causal reasoning in natural, indeterminate situations (Table 16).

Table 16: Applications of a Naturalistic Perspective on Causal Reasoning

Problem	Solution
Confusion with regard to "effect"	Ostensive definitions
Limited data to determine trends	Examine process along with outcomes
Escaping from the "single-cause" mindset	Causal fields
Selecting an appropriate Causal Reasoning format	Explanation Guide
How to shift from understanding to action	Reverse-o-meter

The shift in goals from understanding to action requires a different type of causal reasoning. With the goal of understanding, we prefer accounts that are more complex, such as the one Snook offers (Figure 20). However, if we want to know what to do, we need to prune this diagram. Some of the nodes are just enabling conditions, such as the fall of the Soviet Union. That is not anything we can manipulate. Our strategy here is to use a method we have dubbed a "reverse-o-meter." For each of the nodes or causes, we would ask two questions: how easy will it be to reverse this cause (the criterion of mutability), and how much of an effect will it have? If we limit our scrutiny to only those causes that are easily reversed, and whose reversal will likely reverse the effect, our analysis becomes much more tractable and pragmatic. We have also prepared reverse-o-meter analyses for a health care accident investigation, and for the failure of Xerox Corporation to successfully market personal computers after Xerox PARC pioneered the major features of personal computers.

The loss of common ground due to ambiguity about the "effect" is best combated using ostensive definitions – definitions by example. Thus, if an organization wants to deal with a leadership problem, we would speculate that "leadership problem" means different things to different people. Before investing too much energy into determining the cause of the leadership problem, we would be better off calibrating what "leadership problem" means using specific examples of poor leadership.

One may address the problem of limited trend data (the Sherry Lansing problem discussed earlier), by examining process along with outcome. True, Paramount Studios was doing worse each year, but the sample size was very small. At the same time, Paramount Studios could not wait for 20 years to be sure the trend was reliable. What it could do was investigate the process Sherry Lansing was using to green light movies, to see if it had changed in any important way.

The problem of single-cause mindsets might be offset by a wider use of causal fields – influence diagrams, like the one that Snook presented, along with the simplified version we have offered. We suspect that experienced decision makers are skeptical of single-cause explanations, but they are wary of the complicated diagrams analysts are tempted to produce. Our concept of a reverse-o-meter is a middle ground we think will be more satisfying.

The problem of selecting an appropriate form of explanation is addressed using an explanation guide about the factors affecting preferred formats – perceived audience sophistication, the context (understanding vs. action), the culture, and so forth.

A second class of applications involves the investigation of accidents. In health care, root cause analyses have become popular to ensure that investigators consider a range of factors, and not just the "blunt end" – the person, usually a nurse, who made the final error in the sequence. Reason's distinction between blunt end and sharp end contributors has sensitized investigators to factors such as fatigue, lack of training, competing pressures, and so forth. We collaborated with Nicholas Sevdalis, Dept of Biosurgery and Surgical Technology at the Imperial College in London to review the London protocol for incident review. We made a number of suggestions about how to improve this process, primarily suggesting the value of the reverse-o-meter to focus the analysis and to particularize it so that not all analyses come up with the same recommendations. We suspect that one of the weaknesses of Reason's Swiss cheese model of accidents is that it encourages organizations to try to eliminate all possible error paths. We believe that is unrealistic and unproductive. Flaws can exist for many years until the fateful convergence occurs. Flaws are most apparent in hindsight, and the effort to eliminate potential flaws may be counterproductive in complex situations – better to prepare the organizations to adapt quickly, as suggested by the proponents of resilience engineering.

A third class of applications is to provide training in causal reasoning for indeterminate situations. During Phase II we offered a one-day causal reasoning working for Defence Science and Technology Agency in Singapore, as a pilot for such training.

REFERENCES

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to Cognitive Task Analysis*. Cambridge, MA: The MIT Press.

Dörner, D. (1989/1996). *The logic of failure: Why things go wrong and what we can do to make them right*. Reading, MA: Perseus.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99*, 3-19.

Evens, M., & Michael, J. (2006). *One-on-one tutoring by humans and computers*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Feltovich, P., Hoffman, R., Woods, D., & Roesler, A. (2004). Keeping it too simple: How the reductive tendency affects cognitive engineering. *IEEE Intelligent Systems, 19*(3), 90-94.

Fugelsang, J. A., Thompson, V. A., & Dunbar, K. N. (2006). Examining the representation of causal knowledge. *Thinking and Reasoning, 12*(1), 1-30.

Gopnik, A., & Schulz, L. (Eds.). (2007). *Causal learning: Psychology, philosophy, and computation*. New York: Oxford University Press.

Hume, D. (1739-1740). *A treatise of human nature*. London: Anonymous. Reprinted by New Vision Publications, Sioux Falls, SD.

Kahneman, D., & Varey, C. A. (1990). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology, 59*, 1101-1110.

Klein, G., Ross, K. G., Moon, B. M., Klein, D. E., Hoffman, R. R., & Hollnagel, E. (2003). Macrocognition. *IEEE Intelligent Systems, 18*(3), 81-85.

Litman, J. A. (2008). Interest and deprivation dimensions of epistemic curiosity. *Personality and Individual Differences, 44*, 1585-1595.

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive* (2002 ed.). Honolulu, HI: University Press of the Pacific.

Mlodinow, L. (2008). *The drunkard's walk: How randomness rules our lives*. NY: Pantheon Books.

Parker, M. (2008). *Panama Fever: The epic story of one of the greatest human achievements of all time -- the building of the Panama Canal*. New York: Doubleday.

Reason, J. (1990). *Human error*. Cambridge, MA: Cambridge University Press.

Sloman, S. (2005). *Causal models. How people think about the world and its alternatives*. New York: Oxford University Press.

Snook, S. A. (2000). *Friendly fire: The accidental shootdown of U.S. Black Hawks over northern Iraq*. Princeton, NJ: Princeton University Press.

Tognazzini, B. (1998). Designing single-use applications. from http://www.asktog.com/columns/010minaturegolf.html.

Tufte, E. (2001). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Watson, J. D., & Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature, April 25*, 737-738.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*, 1049-1062.

APPENDIX A: Causal Reasoning Scenarios

Economic Collapse

On January 1, 2007 the U.S. economy was looking very strong. The stock market was going up. The Dow Jones Index, a measure of consumer confidence in the U.S. economy closed at 12,474 on January 3, 2007. It went up and up, closing at 14,164 on October 9, 2007. Then it came crashing down, closing at 8,776 on December 31, 2008. It eventually bottomed out at 6,547 on March 9, 2009, less than half its value at its peak. Banks failed, investment houses closed, lending all but ceased, housing prices dropped dramatically, and a massive government bail-out was needed to keep the U.S. economy going.

Success of the Surge in Iraq

The recent success of the U.S. – led coalition in Iraq. After the invasion of Iraq in 2003, it appeared that the U.S.-led coalition was bogged down in an unwinnable insurgency conflict, eerily similar to the Vietnam War that the U.S. fought. But then in 2007, things turned around, and Iraq seems to be in much better shape. There are still horrifying terror attacks, but these occur much less often, and fewer civilian casualties are occurring.

APPENDIX B: List of Explanations

Explanations for Economic Collapse Scenario

Explanation #1: Alan Greenspan, the head of the U.S. Federal Reserve, tried to boost prosperity by keeping interest rates low but this backfired when it led many people to buy homes they could not afford, creating a housing bubble that ruined the economy.

Explanation #2: Housing prices dropped very sharply in 2007-2008, not leaving much time for homeowners or investors to cope, and this caused the U.S. economy to collapse.

Explanation #3: The pressure for growth and profits pushed the economy to unhealthy and unsustainable levels.

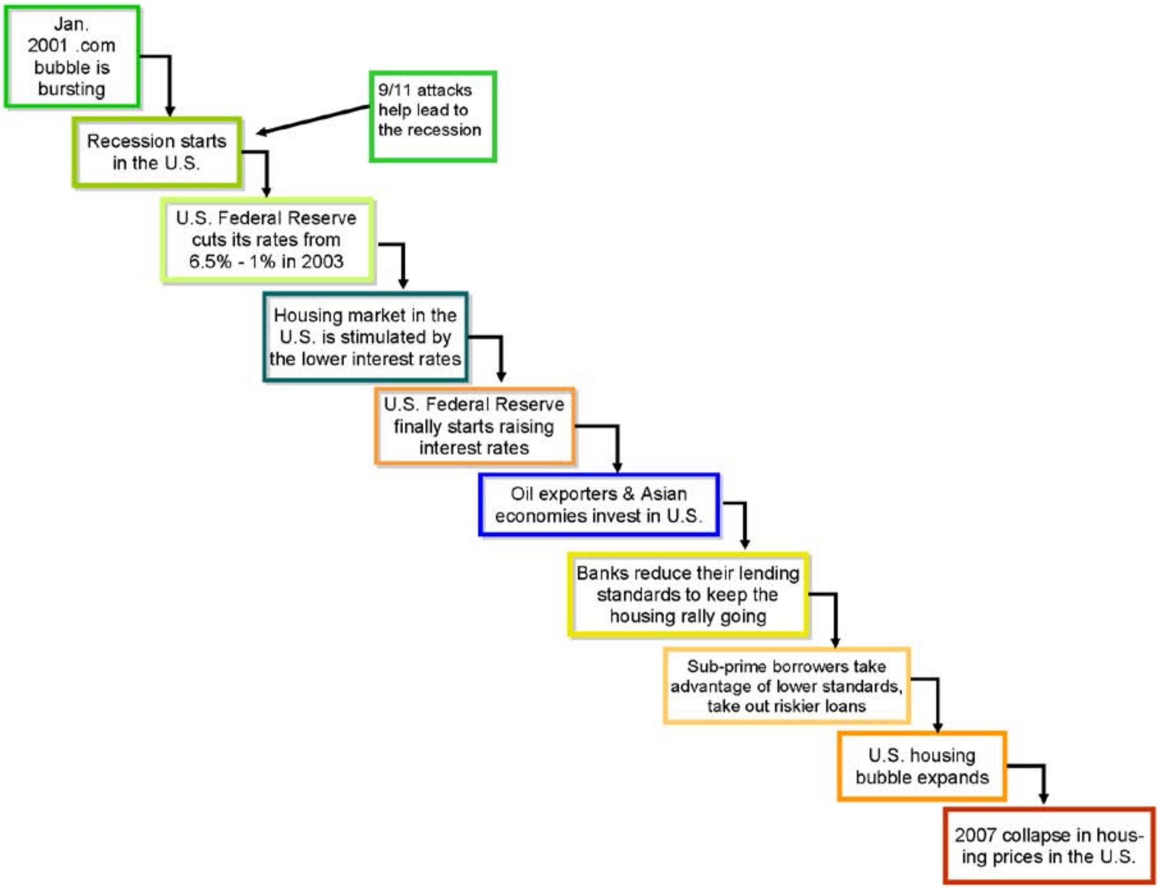
Explanation #4: It all boils down to "greed." Everyone, from investors to homeowners to stock brokers, was blinded by greed that overcame good judgment.

Explanation #5: It was a combination of all the things listed below:

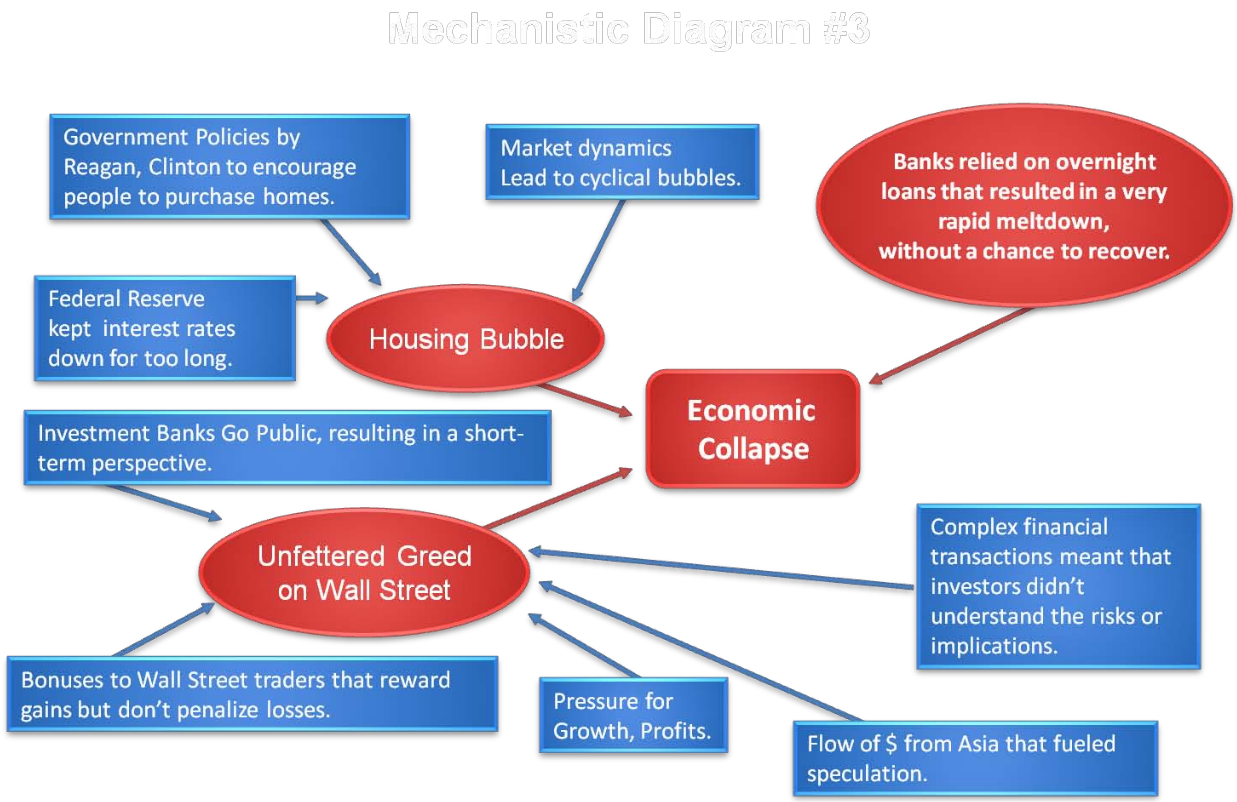
- "Good times," and the belief they would continue
- Alan Greenspan's decision to continue to cut interest rates
- Inadequate regulations
- Wall Street
- Homeownership obsession
- Too much money flowing into U.S. from abroad
- Myth of the rational market
- Misplaced reliance on risk analysis methods.
- U.S. citizens have high debt and little savings
- George W. Bush
- Bill Clinton signing the Commodity Futures Act, and repeal of Glass-Steagall Act
- Rating agencies
- Decision to let Lehman Brothers fail

Explanation #6:

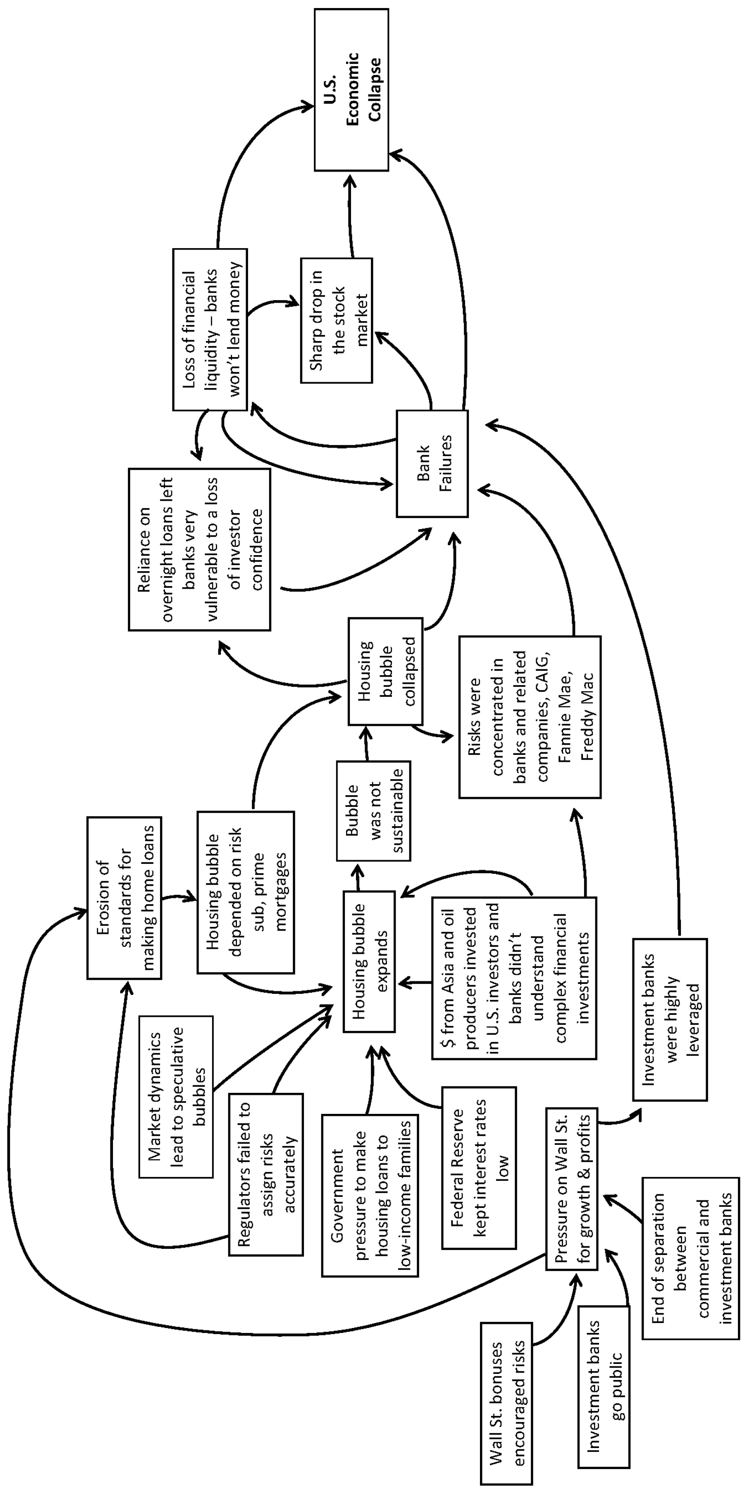
It was like a chain reaction, as shown in the diagram below.



Explanation #7: It was a few primary forces that intersected, as shown in the diagram below.



Explanation #8: There were many forces that interacted in complex ways, as shown in the diagram below.



Explanations for Iraq Scenario

Explanation #1: The U.S. decided to launch a surge – to increase the number of soldiers in order to turn the tide of the conflict against the Iraqi insurgents.

Explanation #2: In June, 2006, the U.S. managed to kill Abu Musab al-Zarqawi, the head of AQI, and the mastermind behind hundreds of bombings, kidnappings and beheadings. This demoralized the AQI terrorists, deprived them of leadership, and gave the Iraqi people more faith in the U.S.

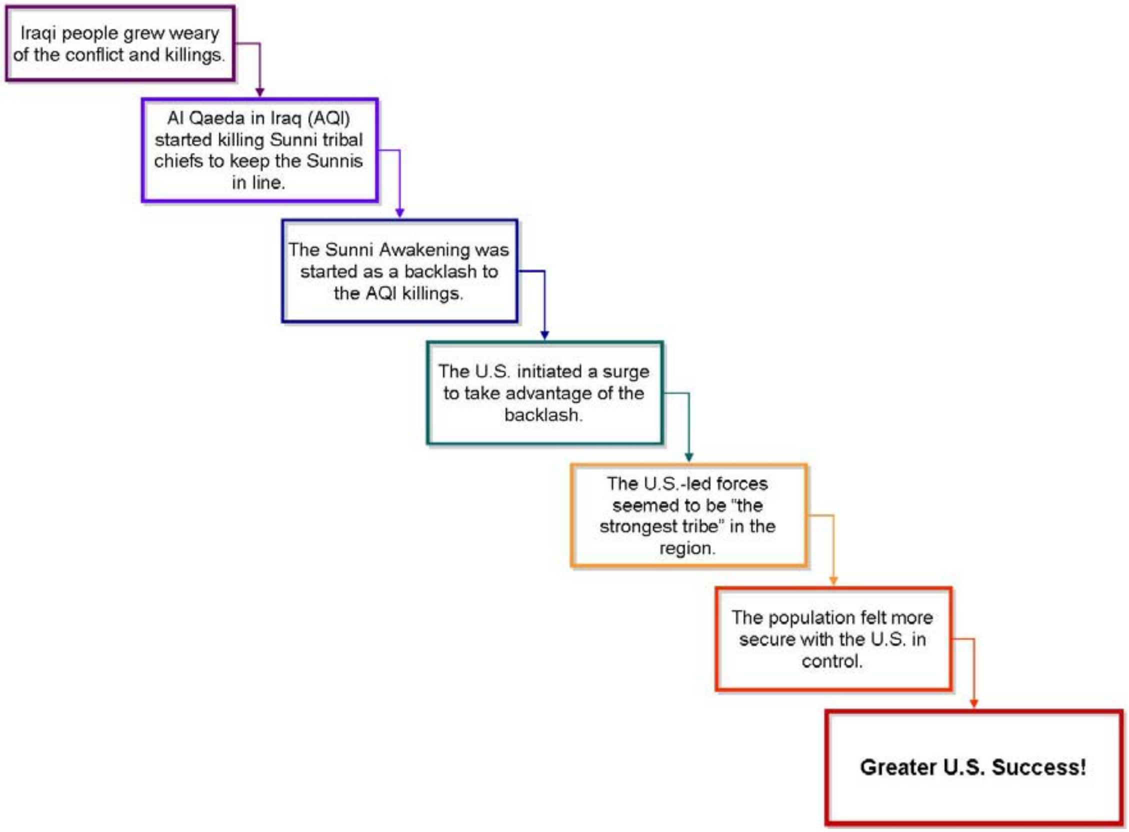
Explanation #3: The insurgent attacks on innocent civilians turned the population against the terrorists.

Explanation #4: The U.S. came to be seen as "the strongest tribe," and the Iraqis accepted the futility of supporting the forces of insurgency.

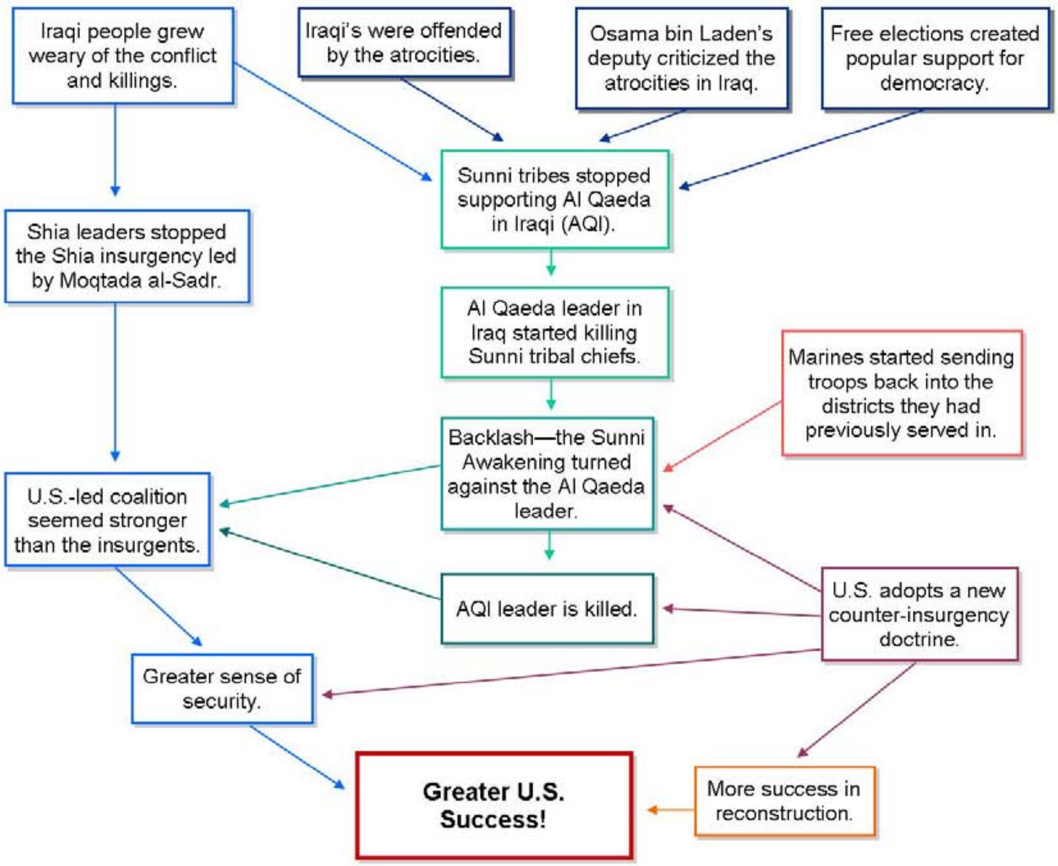
Explanation #5: It was a combination of all the things listed below:

- The surge of U.S. soldiers in 2007.
- The death of the terrorist leader al-Zarqawi.
- Disapproval of the atrocities committed by the terrorists and insurgents.
- The critique of the atrocities issued by the al Qaeda second-in-command Ayman al-Zawahiri, from Afghanistan/Pakistan, complaining that AQI was acting inappropriately and turning the populace against them.
- The policy of AQI to kill Iraqi Sunni tribal leaders to frighten the Sunni tribes into supporting them.
- The initiation of the "Sunni Awakening" to resist AQI by allying with the U.S. forces in Anbar province.
- The new procedure from the U.S. Marine Corps to rotate troops back into the same area that they had previously served on their last rotation, in order to build on the relationships they had forged with locals.
- The new counter-insurgency doctrine the U.S. adopted as a change in strategy and tactics.
- The perception that the U.S. forces were stronger than the insurgents, and had become "the strongest tribe," which encouraged Iraqis to change their alliance.
- The perception that the U.S. was making Iraqis more secure, and that without the U.S. Iraq might plunge into a civil war of Shia versus Sunni.
- The coalition reconstruction efforts were finally paying off.
- The free elections that demonstrated to Iraqis that democracy might work. Even though the candidate that the U.S. preferred was beaten, the U.S. did not overturn the results.
- Moqtada al-Sadr muzzled his Shia militia, after the Shia-dominated government in Iraq defeated his forces.
- The war-weariness felt by Iraqis, particularly in the aftermath of the near civil war between Shia and Sunni.

Explanation #6: It was like a chain reaction, as shown in the diagram below.



Explanation #7: It was a few primary forces that intersected, as shown in the diagram below.



LIST OF ACRONYMS

AF	Air Force
AFRL	Air Force Research Laboratory
AIDS	Acquired Immune Deficiency Syndrome
AQI	Al-Qaeda in Iraq
AWACS	Airborne Warning and Control Systems
DNA	Deoxyribonucleic Acid
EUG	Experiential User Guide
FAQ	Frequently Asked Question
GPS	Global Positioning System
HIV	Human Immunodeficiency Virus
IFF	Identification Friend or Foe
JCAT	Java Causal Analysis Toolkit
MBA	Master of Business Administration
NFC	Need for Cognition
NOEM	National Operational Environmental Model
PERT	Program Evaluation and Review Technique
RHXS	Sensemaking and Organizational Effectiveness Branch
ROTC	Reserve Officer' Training Corps
TACS	Technology for Agile Combat Support
SME	Subject-Matter Expert
U.S.	United States
U.S.S.R.	Union of Soviet Socialist Republics